Health Serv Outcomes Res Method (2009) 9:69–83
DOI 10.1007/s10742-009-0047-1

Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities

Marc N. Elliott · Peter A. Morrison · Allen Fremont · Daniel F. McCaffrey · Philip Pantoja · Nicole Lurie

Received: 30 September 2008 / Revised: 7 January 2009 / Accepted: 24 March 2009 /
Published online: 10 April 2009

Abstract Commercial health plans need member racial/ethnic information to address disparities, but often lack it. We incorporate the U.S. Census Bureau's latest surname list into a previous Bayesian method that integrates surname and geocoded information to better impute self-reported race/ethnicity. We validate this approach with data from 1,921,133 enrollees of a national health plan. Overall, the new approach correlated highly with self-reported race-ethnicity (0.76), which is 19% more efficient than its predecessor (and 41% and 108% more efficient than single-source surname and address methods, respectively, $P < 0.05$ for all). The new approach has an overall concordance statistic (area under the Receiver Operating Curve or ROC) of 0.93. The largest improvements were in areas where prior performance was weakest (for Blacks and Asians). The new Census surname list accounts for about three-fourths of the variance explained in the new estimates. Imputing Native American and multiracial identities from surname and residence remains challenging.

M. N. Elliott (✉) · A. Fremont · P. Pantoja
RAND Corporation, 1776 Main Street, Santa Monica, CA 90407, USA
e-mail: elliott@rand.org

A. Fremont
e-mail: fremont@rand.org

P. Pantoja
e-mail: pantoja@rand.org

P. A. Morrison
RAND Corporation, 3 Eat Fire Springs Road, Nantucket, MA 02554, USA
e-mail: peterm3636@aol.com

D. F. McCaffrey
RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213, USA
e-mail: danielm@rand.org

N. Lurie
RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202, USA
e-mail: lurie@rand.org

Keywords Bayesian inference · Health disparities · Race and ethnicity · Health insurance

1 Introduction

Efforts to measure racial/ethnic disparities in health care have been frustrated by limited availability of racial/ethnic data, particularly for enrollees of commercial health plans (Elliott et al. 2008b; Fremont and Lurie 2004). Collection of such data has proceeded slowly despite recommendations by the Institute of Medicine and the National Academy of Sciences that self-reported race/ethnicity data be voluntarily obtained from individuals (Institute of Medicine 2002; National Research Council 2004). Efforts to collect such data by self-report are under way in several states: Massachusetts health care reform legislation now requires collection of race/ethnicity from all hospitalized patients (Boston Public Health Commission 2006); California SB 853 and related regulations now require HMO plans to collect race/ethnicity information (California State Senate 2007). Plans participating in the National Health Plan Collaborative (NHPC) to Improve Quality and Eliminate Disparities also voluntarily collect their enrollees' self-reported race/ethnicity (National Health Plan Collaborative 2006). However, even plans that have made major efforts to collect such self-reported data over a period of several years report that it is challenging to obtain self-reported information on more than about a third of enrollees.

1.1 Surname and geocoding approaches

The absence of such data has spurred development of methods to estimate an individual's race/ethnicity indirectly from other sources. Two such methods are geocoding and surname analysis.

Geocoding links an individual's address to a census measure of their neighborhood's racial/ethnic population makeup and uses that measure as a basis for inferring the individual's race/ethnicity. Blacks are racially concentrated in many neighborhoods around the country (Fremont et al. 2005); in those areas, geocoding alone can be fairly effective in distinguishing Blacks from Whites. Because Hispanics, Asians, and many Native Americans tend to live in far less segregated neighborhoods than Blacks (Logan 2001; Massey and Denton 1989), geocoding alone cannot reliably identify members of these minority groups. An alternative approach is simply to use the racial/ethnic prevalences shown in the Census 2000 SF1 block group data as the probabilities of an individual belonging to each of the major racial/ethnic groups. This latter approach, hereafter "Geocoding Only (GO)," more fully incorporates the available census information than a "cut point" approach and provides information for several multiple racial/ethnic groups (see Elliott et al. 2008b).

Surname analysis infers race/ethnicity from surnames (last names) that are distinctive to particular racial/ethnic groups. Initially, surname analysis entailed using dichotomous dictionaries to identify Hispanics and various Asian nationalities (Abrahamse et al. 1994; Falkenstein 2002; Kestenbaum et al. 2000; Lauderdale and Kestenbaum 2000; Perkins 1993). Construction of these lists emphasized high specificity (i.e., persons whose surnames appear on the list have a high probability of self-reported Hispanic ethnicity or Asian race, respectively). Although useful, these lists do not fully utilize all the information surnames might convey regarding race/ethnicity. First, no distinction is made

among the more informative and less informative individual surnames on a list. Second, the lists exclude surnames with only intermediate specificity, which still distinguish race/ethnicity far better than chance.

1.2 Previous hybrid approaches

Both GO and surname analysis methods used alone have notable limitations: standard surnames lists cannot distinguish Blacks from non-Hispanic Whites; geocoding is of limited use in identifying Hispanics or Asians. Hybrid methods designed to combine both methods, such as the *Categorical Surname and Geocoding* approach (CSG) (Fiscella and Fremont 2006), can distinguish Blacks, Asians, and Hispanics from non-Hispanic Whites. A subsequent Bayesian hybrid (Elliott et al. 2008b), the *Bayesian Surname and Geocoding* method (BSG), made substantial improvements on the CSG, yielding 21% more efficient use of data (121% the relative efficiency of CSG) for identifying individual race/ethnicity than CSG (with the greatest gains for Blacks) and 74% more efficient use than geocoding only (GO), where relative efficiency is defined as the average ratio of squared correlations between estimated and self-reported racial/ethnic indicators.

The Bayesian approach is analogized from medical diagnostic testing, where the (posterior) probability of having a disease depends upon (1) an individual's "prior" probability of having the disease (e.g., the base rate for the individual's risk group) and (2) the outcome of a diagnostic test. Bayes' Theorem updates prior probabilities with test results by considering the *sensitivity* and *specificity* of the diagnostic test to produce an updated (posterior) probability, the *positive predictive value*.

More general forms of Bayes' Theorem allow for tests with more than two outcomes. The BSG approach treated the racial/ethnic distribution of the Census 2000 block group where an individual lives as a four-category prior (either Hispanic, African-American, Asian, or non-Hispanic White/Other). It then used (1) the combined *results* of the Census Bureau Spanish Surname List and the Lauderdale-Kestenbaum Asian Surname List as a diagnostic test with three possible outcomes (surname appears on Asian list regardless of appearance on Spanish Surname list, surname appears on Spanish but not Asian list, surname appears on neither surname list); and (2) the *sensitivity and specificity* of these lists to *update* the prior probabilities of membership in each of the four racial/ethnic categories. This results in a final (posterior) set of four probabilities of membership in the four racial/ethnic groups for each individual.

1.3 An improved surname list and an improved Bayesian approach

In 2007, the U.S. Census Bureau released a national tabulation of unprecedented detail, showing surnames classified by self-reported race/ethnicity, based on almost 270 million individuals with valid surnames enumerated on Census 2000 (see Word et al. (2008) for a description and http://www.census.gov/genealogy/www/freqnames2k.html for a machine-readable Excel file of data). Of six million unique surnames, just 275 accounted for 26% of all individuals; conversely, fully 5 million other surnames each were listed by four or fewer individuals. The 151,671 surnames listed by 100 or more individuals, along with each surname's self-reported racial/ethnic distribution, are publicly available and represent 89.8% of all individuals enumerated on Census 2000. In accordance with Office of Management and Budget standards (U.S. Office of Management of Budget 1997), the Census elicits a respondent's Hispanic ethnicity, then asks the respondent to self-identify

with one or more of the following six races: White, Black, American Indian/Alaska Native (AI/AN), Asian, Native Hawaiian/Other Pacific Islander (PI), and Some Other Race. "Some Other Race" is endorsed by about 7% of individuals, primarily individuals of Hispanic ethnicity (U.S. Office of Management of Budget 1997) seeking to indicate national origin. Word et al. (2008) used standard Census procedures (Schenker and Parker 2003) to delete and reimpute race for these individuals; classified all of those who endorsed Hispanic ethnicity as Hispanic; combined Asian and Pacific Islander categories; and classified non-Hispanic individuals who chose more than one race (after these prior steps) as "multiracial." Thus, for each surname with 100 or more occurrences nationally, the publicly available tabulation shows the frequency of occurrence in each of six mutually exclusive categories—(1) Hispanic, (2) White, (3) Black, (4) Asian and Pacific Islander (API), (5) American Indian/Alaska Native (AL/AN), and (6) Multiracial. To preserve confidentiality, exact counts for specific racial/ethnic groups were suppressed for a given surname when any of these six categories had fewer than five occurrences; and when only a single category had fewer than five occurrences for a given surname, its count and the category with the second fewest occurrences were suppressed.

We incorporated this additional information into the general framework of the previous Bayesian algorithm in a new approach, referred to hereafter as the *Bayesian Improved Surname Geocoding* method (BISG). Our focus below is a comparison of the accuracy of this newly updated Bayesian approach to that of the earlier Bayesian approach (BSG). We also compare the new method's accuracy to that of geocoding only (GO) and the new surname list only (NSO) in order to distinguish the value of the enhanced surname list from other sources of information for inferring race/ethnicity.

2 Method

Table 1 provides an overview of the methods reviewed in this paper.

The new BISG method relies on new data combined with Bayesian methods for inferring race/ethnicity. We describe the data, then our methods for determining race/ethnicity, and finally our method for assessing the accuracy of BISG relative to the earlier alternatives.

Table 1 Summary of four methods compared (GO, NSO, BSG, BISG)

Method	Prior probabilities	Test/updating of probabilities	Output
GO	4-category census block group race/ethnicity proportions	None	4-category probability of race/ethnicity
NSO	6-category race/ethnicity proportions from surname (new census list)	None	6-category probability of race/ethnicity
BSG	4-category census block group race/ethnicity proportions	Membership on older Spanish and Asian surname lists uses as test for Bayesian updating	4-category probability of race/ethnicity
BISG	6-category race/ethnicity proportions from surname (new census list)	6-category census block group race/ethnicity proportions relative to U.S. as a whole used as "test" for updating	6-category probability of race/ethnicity

2.1 Data for validation

We used 2006 national enrollment data from Aetna, a large national health plan. The data set consists of self-reported race/ethnicity, surname, address of residence, and gender for the 1,921,133 enrollees who voluntarily provided this information to the plan for quality monitoring and improvement purposes (30% of all enrollees). Surname and address (linked to block groups in the US 2000 Census SF1 file) were used as inputs in the estimation algorithms. Self-reported race/ethnicity was used only to evaluate the algorithms' performance after they had estimated the race/ethnicity of individuals.

While voluntarily reported race/ethnicity was predominantly non-Hispanic White (77.0%), the data set included a reasonable distribution of Hispanics (9.2%), Blacks (8.2%), and Asians/Pacific Islanders (5.1%), with 0.3% American Indian/Alaska Native (AI/AN) and 0.2% multiracial; 51.2% were female. As is typical of commercially insured populations, Hispanics and Blacks were somewhat underrepresented, and Asians and Whites somewhat overrepresented, relative to the U.S. population as a whole. This project was reviewed by the RAND IRB and all data disclosed to the authors by health plans were in compliance with HIPAA regulations.

2.2 Implementation of the BISG

2.2.1 Refinement of new surname probabilities

The new surname list has limitations: (1) the suppression of exact counts for surnames with infrequent occurrences in some groups and (2) the omission of surnames with fewer than 100 occurrences.

When a surname had more than 100 occurrences but fewer than five of them fell in at least one racial/ethnic category, exact counts were unavailable for $k = 2$–5 racial/ethnic categories for a given surname, but both k and n, the sum of the occurrences in the missing categories, were known (because total occurrences for the surname are provided).[1] We imputed the suppressed counts for the missing cells as n/k in such instances.[2] This resulted in a vector of six predicted probabilities for each of 151,671 names with 100 or more occurrences, covering 89.8% of the general U.S. population (and 89.4% of our sample population).

In order to infer the race/ethnicity of the one-tenth of the population with unlisted names (i.e., names with fewer than 100 total occurrences), we subtracted the racial/ethnic counts for listed surnames (i.e., names with 100 or more occurrences) on the US Census List from racial/ethnic counts calculated from 2000 Census SF1 counts for *all* individuals.[3] The remaining counts in each racial/ethnic category describe the racial/ethnic distribution of the "unlisted name" population (i.e., all persons with surnames occurring fewer than 100 times each on Census 2000): 70.5% White, 11.1% Hispanic, 11.3% Black, 7.0% API, 0.8% multiracial, and 0.9% AI/AN (Jirousek and Preucil 1995). This distribution compares to 69.5% White, 12.5% Hispanic, 12.2% Black, 3.8% API, 1.2% multiracial, and 0.7% AI/AN

[1] The specific counts that were suppressed were also known.

[2] Exploratory analyses (not shown) demonstrated better overall predictive performance with this approach than with several alternatives we considered.

[3] Because the 2000 Census SF1 file includes an "other race" category not used in the Census surname list, we reassigned responses at the level of the block group using Iterative Proportional Fitting (Jirousek and Preucil 1995), an approach similar to that used by Word et al. (2008).

for the overall Census 2000 population. Thus, those with low-frequency names (<100 occurrences) correspond closely to the entire US population, with the exception that they were more likely to be Asians or Pacific Islanders.

2.2.2 *BISG Bayesian updating formulas*

We then used the vectors of six racial/ethnic probabilities for each listed surname (corrected for suppression and for low-frequency surnames as described above) as input to the new BISG algorithm. This algorithm updates these "prior probabilities" with geocoded block group proportions for these same six groups from the (reassigned) 2000 Census SF1 file to generate posterior probabilities. The Bayesian calculations are as follows:

Let J equal 151,672, the number of names on the enhanced surname list plus one to account for names not on the list and let K equal 208,125, the number of block groups in the 2000 census with any population. We define the prior probability of a person's race on the basis of surname,[4] so that for a person with surname $j = 1,\ldots, J$ on the list, the prior probability for race, $i = 1,\ldots,6$, is $p(i|j)$ = proportion of all people with surname j who report being of race i in the enhanced surname file (*the probability of a selected race given surname*).

We then update this probability on the basis of Census block group residence. For block group $k = 1,\ldots,K$, $r(k|i)$ = proportion of all people in redistributed SF1 file who self report being race i who reside in Census Block Group k (*the probability of a selected Block Group of residence given race/ethnicity*).

We require an assumption that the probability of residing in a given Block Group, given a person's race, does not vary by surname.

Let $u(i, j, k) = p(i|j) \times r(k|i)$ then $q(i|j, k)$, the updated (posterior) probability of being of race/ethnicity i given surname j and census block group of residence k, can be calculated as follows, according to Bayes' Theorem and the above assumption.

$$q(i|j,k) = \frac{u(i,j,k)}{u(1,j,k) + u(2,j,k) + u(3,j,k) + u(4,j,k) + u(5,j,k) + u(6,j,k)}$$

Note that all parameters needed for BISG posterior probabilities are derived only from Census 2000 data, and that none are derived from health plan data or other administrative sources. As such, BISG performance estimates in predicting race/ethnicity within health plans do not reflect overfitting and are not subject to "shrinkage" in applications to other data sets.

2.3 Summary comparison of the four algorithms

Next, we compare the performance of BISG, BSG, GO, and NSO. Table 1 compares these four approaches in terms of their inputs, updating, and outputs. Because BSG and GO produce a vector of four probabilities (Hispanic, Black, API, and White/Other), our primary comparisons combine three of the six categories from NSO and BISG (AI/AN, multiracial, and White) into a single White/Other category for comparability with GO and BSG. Secondary analyses for NSO and BISG employ the full six categories, noting that performance differs only among the three categories subject to this recoding.

[4] We present the results treating surname information as the prior that is updated by the geocoded information; however, we would obtain the same results if we treated the geocoded information as the prior and updated with the surname data.

2.4 Evaluation

We compare the four approaches in terms of how closely the estimates of race/ethnicity that they produce match those derived from self-reported race/ethnicity information for the same individuals using the Aetna validation data. We describe a performance metric applicable to all four approaches. We then compare the relative efficiency of the four methods according to this metric—accuracy of predicting individual race/ethnicity—the extent to which those assigned higher probabilities of a given race/ethnicity are more likely to self-report that race/ethnicity.

Following Elliott et al. (2008b), we define the efficiency of prediction for a given racial/ethnic category as the squared correlation between the predicted probability for a given racial/ethnic category and the corresponding dichotomous indicator of true self-reported race/ethnicity in the Aetna data (Black/not Black, Asian/not Asian, etc.).[5] For each method, we summarize performance across all four racial/ethnic categories with the average of the six squared correlations weighted by the prevalence on each racial/ethnic group. McCaffrey and Elliott (2008) establish that the efficiency of an analysis that directly uses predicted racial/ethnic probabilities rather than true race/ethnicity indicators to estimate racial/ethnic disparities on an outcome of interest in a regression model is well approximated by the above definition of efficiency (the squared correlation between estimated and true race/ethnicity).[6] Hence, we use the ratio of squared correlations for pairs of methods to measure their *relative efficiency* in predicting each individual racial/ethnic category. We summarize the overall performance through a weighted average of the six squared correlations for individual racial/ethnic groups, weighting by the self-reported population proportions. To say that method A has a relative efficiency of 300% when compared to method B means that the accuracy of an analysis testing for difference in an outcome of interest among racial/ethnic groups using method A with a sample of a given size is the same as what would be obtained with three times the sample size using method B. In this instance, we would also say that method A is (300–100%) = 200% more efficient than method B.

We provide additional information about the performance of the BISG. First, we show the distribution of BISG predicted probabilities for our 2006 Aetna data. Second, we analyze BISG correlation with self-reported race/ethnicity separately by gender. Finally, we calculate the concordance statistic (Hanley and McNeil 1982), or area under the curve, to summarize the BISG's performance with the Aetna data.

3 Results

3.1 Predicting individual race/ethnicity: comparing BISG, BSG, NSO, and GO

Table 2 displays the correlation of predicted race/ethnicity with self-reported race/ethnicity for each of the four methods and four racial/ethnic groups in the primary data set. All reported correlations are statistically significant and differ across methods at $P < 0.05$.

[5] Because the racial/ethnic categories are mutually exclusive, estimates for the groups are negatively correlated.

[6] A squared correlation of 0.49 between estimated race/ethnicity and self-reported implies approximately 49% efficiency relative to known race/ethnicity for estimating a disparity between two racial/ethnic groups under the assumptions in that paper.

Table 2 Correlation of individual predicted race/ethnicity with self-reported race/ethnicity ($n = 1{,}921{,}133$)

	Correlation with self-reported race/ethnicity				Weighted average
	Hispanic	Asian	Black	White/Other	
BISG	0.82	0.77	0.70	0.76	0.76
BSG	0.80	0.69	0.62	0.72	0.70
GO	0.49	0.34	0.57	0.55	0.53
NSO	0.79	0.74	0.40	0.64	0.64

All differences in correlations by methods are significant at $P < 0.05$

BISG predictions correlate with individual indicators of race/ethnicity at 0.70–0.82, with a weighted average correlation of 0.76. This represents an increase in efficiency over BSG that averages 19% (119% the relative efficiency of BSG), ranging from 4% for Hispanics to 25% for Asians and 27% for Blacks. Notably, the biggest improvement was where BSG was weakest (Blacks) and the smallest improvement was where BSG was strongest (Hispanics). BISG thus not only improved the performance of the BSG, but also made performance less disparate across racial/ethnic groups, although performance is still highest for Hispanics (0.82 correlation with self-report) and lowest for Blacks (0.70 correlation with self-report).

Overall, BISG is 108% more efficient than GO and 41% more efficient than NSO, suggesting a larger role for the new surnames than for census block group of residence. As expected, since it combines the two sources of information, BISG exceeds the predictive power of both NSO and GO for each racial/ethnic group, with NSO coming fairly close for Hispanic and Asian individuals. Table 3 partitions the total predictive power of BISG (variance of dichotomous self-reported indicators explained) into three sources—jointly determined by residential location (GO) and surnames (NSO), unique to location, and unique to surnames, using squared correlations from the GO, NSO, and BISG models. If a, b, and c are the correlations of GO, NSO, and BISG with self-report, respectively, $(c^2 - b^2)/c^2$ represents the proportion of the squared BISG variance uniquely explained by GO, $(c^2 - a^2)/c^2$ represents the proportion uniquely explained by NSO, and $(a^2 + b^2 - c^2)/c^2$ represents the proportion jointly explained.

As can be seen, half of the total predictive power of BISG is unique to surnames (using the new surname list), about a quarter is unique to location, and about a quarter is jointly determined (could have been identified by either surname or residential location). As expected, these proportions vary strongly by race/ethnicity, with surnames alone responsible for 81% of BISG's predictive power for Asians and 64% of BISG's predictive power for Hispanics, but only 33% of BISG's predictive power for Blacks.

Table 3 Proportion of BISG information explained by new surnames, geocoding ($n = 1{,}921{,}133$)

	Proportion of BISG information explained				Weighted average
	Hispanic	Asian	Black	White/Other	
Jointly determined	0.31	0.13	0.01	0.23	0.24
Unique to GO	0.05	0.07	0.66	0.29	0.26
Unique to NSO	0.64	0.81	0.33	0.48	0.50

Table 4 Correlation of individual predicted race/ethnicity with self-reported race/ethnicity ($n = 1{,}921{,}133$)

	Hispanic	Asian	Black	AI/AN	Multiracial	White	Weighted average
BISG	0.82	0.77	0.70	0.11	0.02	0.76	0.75
NSO	0.79	0.74	0.40	0.08	0.01	0.64	0.64

All differences in correlations by methods are significant at $P < 0.05$

Table 4 is similar to Table 2, but breaks down White/Other into White, AI/AN, and multiracial; it also is restricted to BISG and NSO, the only methods applicable to six-category race/ethnicity. By definition, performance is unchanged for Hispanic, Asian, and Black individuals. Performance for Whites here is essentially the same as for Whites/Others in the four-category predictions. Although correlations exceed chance ($P < 0.05$ in all cases), BISG prediction of AI/AN is poor (correlation of 0.11 with self-report) and BISG prediction of multiracial is very poor (correlation of 0.01 with self-report).

3.2 Predicting individual race/ethnicity: additional evaluation of BISG

Table 5 shows the distribution of BISG probabilities for the six racial/ethnic categories in the Aetna data. BISG probabilities of Hispanic, Asian, and Black have a bimodal distribution, with 83.1–93.0% of probabilities below 0.05 and 2.5–2.9% of probabilities 0.90 or higher. BISG probabilities of White have a bimodal distribution that roughly mirrors that of these three groups with 68.3% of probabilities 0.90 or higher and 11.2% of probabilities from 0.05 to 0.20. Less than 0.1% of cases have BISG probabilities of AI/AN or Multiracial exceeding 0.05.

Table 6 summarizes correlation of BISG predictions with self-reported race/ethnicity by gender. For Hispanic and Asian categories, where surname lists are most important (see Table 3), performance for males is moderately higher than for females, with relative efficiencies that are 13% and 11% higher, respectively. Higher performance for males is

Table 5 Percentage of individuals with specified BISG probabilities, for each of the six predicted racial/ethnic categories

Bayesian probability	Hispanic	Asian	Black	AI/AN	Multiracial	White
0 to <0.05	87.9	93.0	83.1	100.0	100.0	4.3
0.05 to <0.20	2.5	2.6	8.6	0.0	0.0	11.2
0.20 to <0.50	0.9	0.7	2.6	0.0	0.0	3.6
0.50 to <0.90	6.2	0.7	2.9	0.0	0.0	12.6
0.90 to 1	2.5	2.9	2.8	0.0	0.0	68.3

Table 6 Correlations of BISG predictions with self-report, by gender

Gender	Hispanic	Asian	Black	AI/AN	Multiracial	White	Weighted by self-reported frequency
Males	0.84	0.78	0.68	0.11	0.02	0.77	0.77
Females	0.79	0.74	0.71	0.09	0.02	0.74	0.75
M-F	0.05	0.04	−0.03	0.02	0.00	0.03	0.02

All differences in correlations by methods are significant at $P < 0.05$ except for multiracial

more modest for Whites and AI/AN; no gender differences are observed by Multiracial; and for Black, probabilities correspond better to self-report for females than for males. Weighting across all categories by self-report, correlations with self-report are slightly higher for males (0.77) than for females (0.75).

Concordance statistics represent the probability that a randomly selected observation self-reporting a given race/ethnicity would have a higher BISG probability of that race/ethnicity than a randomly selected observation not self-reporting that same race/ethnicity. The concordance statistics were 0.95 for Hispanic, 0.94 for API, 0.93 for Black and White, 0.77 for multiracial, and 0.61 for AI/AN, yielding a weighted average of 0.93 concordance. Concordance statistics represent the area under a Receiver Operating Characteristic (ROC) curve and, when multiplied by two with one subtracted, are equivalent to the Gini coefficient (Hand and Till 2001), which runs from 0 at chance performance to 1 at perfect prediction. The Gini coefficient for Hispanic predictions is thus $0.95 * 2 - 1 = 0.90$, i.e., 90% of the way between chance and perfect performance.

4 Discussion

We have described an improved Bayesian method (the Bayesian Improved Surname Geocoding approach, BISG) for estimating race/ethnicity using a new surname list derived from Census 2000. Applying the Bayes' Theorem to geocoding and surname analysis proves to be an effective means of integrating these two sources of information and substantially improves a previous Bayesian approach (BSG). The BISG discriminates Hispanics, API, Blacks, and Whites well, with concordance statistics of 0.93 or greater for each of these groups.

The advantage of BISG over BSG stems from the use of a surname list of continuous, rather than dichotomous, probabilities. Performance is slightly higher for males than for females for the two categories most reliant on surnames- Hispanic and Asian, which likely reflects more females than males acquiring new surnames less typical of their own race/ethnicity at marriage.

Somewhat surprisingly, the new surname list afforded substantial improvements in the identification of Blacks and non-Hispanic Whites (groups that generally lack distinctive surnames). Among Blacks, there are several common surnames with more than a 50% probability of belonging to Blacks, at rates far above the national prevalence of Blacks; these include Washington (90% Black), Jefferson (75%), Banks (54%), and Jackson (53%) (Word et al. 2008). While these surnames contain substantial information about race, they lack the specificity to make a traditional dichotomous surname list useful. In contrast, there are many surnames (including Yoder, Krueger, Mueller, Koch, Schwartz, and Novak) that are highly specific to Whites, with more than a 97% chance of indicating non-Hispanic White race/ethnicity (Word et al. 2008). Unfortunately, each of these highly specific White surnames is very uncommon, and together they represent but a miniscule fraction of self-identifying non-Hispanic Whites, so a dichotomous surname list would be insufficiently sensitive. An important contribution of the new Census surname list (and of the BISG) is the ability to capture more information from surnames than dichotomous lists allow.

Identifying AI/AN from surname and residence remains difficult at the national level. Surnames contribute almost no information; the most predictive AI/AN surname is Lowery, which indicates only a 4% chance of being AI/AN (Word et al. 2008). Residential location would be highly informative for AI/AN living in several Native American reservation areas in the southwest (Elliott et al. 2008a) but generally uninformative for the

majority of community-dwelling AI/AN when the focus is national. Predicting multiracial endorsement from address and surnames is not currently feasible. While a few names indicate a 15–18% chance of self-identifying as multiracial (e.g., Ali, Khan, and Singh; (Word et al. 2008), such names constitute only a miniscule proportion of those who identify as multiracial.

Beyond its ability to estimate race/ethnicity, the BISG approach has substantial potential for the design and evaluation of interventions to reduce racial/ethnic disparities. Health plans and others could target an intervention at those most likely to belong to a given race/ethnicity, setting the probability threshold at whatever level resources allow. For example, consider a hypothetical intervention aimed at addressing Black-White health disparities among a health plan's members. One could sort members by the predicted probability of being Black and target the 1,000 individuals with the highest probability of being Black; or identify all members whose predicted probability of being Black exceeds 70%. Furthermore, one could proceed the same way within a subset of members (e.g., known diabetics); or to measure disparities in health and health care within a health plan or hospital, tracking trends over time. As noted by Elliott et al. (2008b), it can also be used when estimated race/ethnicity is to be a predictor in multivariate regression for covariate-adjusted measurement of the association of race/ethnicity with any number of outcomes.

One limitation common to all methods of inferring race/ethnicity is that BISG requires somewhat larger sample sizes to estimate disparities than is required with self-reported race/ethnicity, as there is some inherent loss of information relative to a sample of the same size. As a heuristic, imputed race/ethnicity will require about $1/r^2$ times the sample size as would self-reported data for the same accuracy, where r is the correlation of imputed race/ethnicity with self-report, so that BISG requires sample sizes 1.73 times as large as self-report for the same accuracy, given an average $r = 0.76$.

The loss of information from modeling with predicted rather than self-reported race/ethnicity need not result in bias, even though it must result in loss in precision when compared to self-reported racial/ethnic data for the same sample size. For instance, with linear models for the outcomes, modeling with the probabilities of race/ethnicity can yield unbiased estimates of model coefficients, provided the probabilities are unbiased (McCaffrey and Elliott 2008). For nonlinear models, consistent estimates can be obtained by maximum likelihood conditional on the probabilities of racial/ethnic group memberships. Using the probabilities of racial/ethnic group membership to classify individuals into groups and using these classifications for analyses would result in loss of efficiency and biased estimates because of classification errors; such an analytic approach should be avoided.

Although we used data from a large national plan with a diverse membership, results may differ somewhat for those who lack health insurance, are insured by other commercial plans, or are disinclined to report race/ethnicity. To address this concern in part, Table 7 provides a correction equation, based on a multinomial logistic regression (not shown), that accounts for the selection into health insurance observed for this particular national plan. This correction requires the assumption that those who did self-report their race/ethnicity to the plan so far do not differ in race/ethnicity from nonresponders with the same surnames and residential Block Group. As an example, respondent with uncorrected probabilities of 0.923 White, 0.026 Hispanic, 0.018 API, 0.017 Black, 0.015 Multiracial, and 0.001 AI/AN would be updated by Table 5 to values of 0.944 White, 0.015 Hispanic, 0.013 API, 0.024 Black, 0.001 Multiracial, and 0.003 AI/AN. These corrected probabilities recalibrate for the ways in which the observed distribution of self-reported race/ethnicity differ from the overall US population with similar surnames and residential addresses.

Table 7 Formula for correcting BISG for insurance selection, based on multinomial logistic regression predicting self-reported race/ethnicity from BISG predicted probabilities ($n = 1{,}921{,}133$)

To obtain this corrected BISG predicted probability	Start with:	Add BISG predicted probability of Hispanic times this:	Add BISG predicted probability of Asian times this:	Add BISG predicted probability of Black times this:	Add BISG predicted probability of AI/AN times this:	Add BISG predicted probability of Multiracial times this:	To get the sum	The corrected BISG prediction is
Hispanic	−4.5350	**7.1409**	3.3532	1.6087	2.5273	0.8336	x	Exp(x)/(1 + exp(x))
Asian	−4.6648	3.3573	**8.3992**	2.1258	2.4654	4.2897		
Black	−3.9763	2.4090	2.3034	**7.0715**	2.0913	2.8809		
AI/AN	−5.8592	2.0903	2.5493	0.7633	**8.1850**	5.4224		
Multiracial	−7.0198	5.5008	6.7947	6.9253	6.6388	**7.0880**		
White/other	1 − the sum of the other five corrected BISG predictions							

Boldfaced entries indicate that the racial/ethnic categories of the rhos in the columns correspond

If the above assumption holds more strongly than an alternative assumption that health insurance coverage does not differ by race/ethnicity among those with the same surnames and residential block group, corrected estimates may be preferable to uncorrected estimates. If the reverse is true, uncorrected estimates may be preferable. Analyses not shown suggest that corrected and uncorrected estimates are very similar in terms of their correlation with true race/ethnicity (more important for estimating disparities). This is the only equation presented in this paper that is fit to our validation data; as such, the extent to which it generalizes to selection into other commercial health insurance is unknown. When working with a different health plan, one might either apply the Table 7 coefficients (under the assumption that similar patterns of selection into commercial health insurance occur in other national plans) or use multinomial logistic regression to calibrate a selection correction to that plan's self-reported distribution. Future applications to additional health plans will reveal the extent to which patterns of selection are similar across plans.

Our approach requires an assumption that the probability of residing in a given Block Group, given a person's race, does not vary by surname. While one might not expect surnames to vary substantially within race by block group, confidentiality requirements preclude testing this assumption with publicly available data. The dominance of the intended terms in the multinomial logit shown in boldface in Table 7, as well as additional models not shown (which show that unadjusted and multinomial logit adjusted predictions have extremely similar correlations with self-report), provides indirect evidence that this assumption is reasonably satisfied.

While direct use of predicted probabilities is somewhat more complex than using categorical racial/ethnic indicators, Elliott et al. (2008b) provide examples and sample SAS code of how this can be done with BSG, BISG, or other methods that generate predicted probabilities of race/ethnicity. On the other hand, a notable strength of the (simple) BISG is its extreme parsimony and use of readily available data sources. The BISG derives all of its parameters from publicly available Census Bureau data, and the predictive performance illustrated here does not use the validation data to estimate any aspect of the standard BISG predictions. Moreover, the means of combining surname and address information comes directly from Bayes' Theorem, not from fitting to validation data. As such, BISG is not subject to overfitting or shrinkage, as empirically based regression approaches are.

Despite their improvement in accuracy, indirect methods such as BISG cannot replace the information gained from self-reported data. However, they are fast and not resource intensive. Given the multi-year time horizon for health plans (and others) to collect enough information to be actionable, and the strong desire of health plans and others to take action to address racial/ethnic disparities in care, such methods can serve as a bridge until the time when adequate self-reported data are available. Furthermore, when estimating mean values or differences in health measures, self-reported and indirectly estimated racial/ethnic indicators can be combined where both exist in order to smoothly bridge the transition to more complete self-reported data. This could be done in several ways. First, estimates using self-reported and indirect methods could be combined via composite estimation to minimize mean-squared error in the estimates of health measures (see Elliott and Haviland 2007; Ghosh-Dastidar et al. in press). Second, if errors in estimating race/ethnicity are uncorrelated with the health measures, one could integrate self-reported racial/ethnic indicators (as "1"s and "0"s) and indirectly estimated probabilities into a single equation, with a relative efficiency of $z = p + (1 - p)r^2$ when compared to a fully self-reported sample of the same size (where r is the correlation of indirect estimates of race/ethnicity with self-report and p is the proportion of observations for which self-report is available).

We anticipate that new developments will allow even further improvements to this general approach. The use of first name listings may improve prediction estimates, particularly for Blacks, and some Asian subgroups (Morrison et al. 2001), although first names tend to be low in sensitivity. Similarly, future efforts may incorporate American Community Survey estimates to capture more recent demographic information.

Surname analysis and geo-coding offer health services researchers, demographers, health plans, and other stakeholders powerful new ways to identify race/ethnicity from administrative records. BISG extends the power of these methods to a variety of potential applications—for example, to disparities in the work place or access to and enrollment in public programs. Finally, because this information already has geographic links, geographic information systems (GIS) open the door to still greater insights.

Acknowledgments This study was supported, in part, by contract 282-00-0005, Task Order 13 from DHHS: Agency for Healthcare Research and Quality. Additional funding and support was provided by RWJF and the Brookings Institute. Marc Elliott is supported in part by the Centers for Disease Control and Prevention (CDC U48/DP000056). The authors thank Bryan GeoDemographics for their work in modifying SF1 Census files for these purposes and Jacquelyn Chou for assistance with manuscript preparation. We thank plans participating in the National Health Plan Collaborative, particularly Aetna, for sharing selected data to help improve efforts to address disparities in care and improve overall quality.

Disclaimer The contents of the publication are solely the responsibility of the authors and do not necessarily reflect the official views of the DHHS.

References

Abrahamse, A.F., Morrison, P.A., Bolton, N.M.: Surname analysis for estimating local concentration of Hispanics and Asians. Popul. Res. Policy Rev. **13**(4), 383–398 (1994). doi:10.1007/BF01084115

Boston Public Health Commission: Data Collection Regulation. Boston, MA (2006)

California State Senate: Senate Bill Analysis of SB 853. Sacramento, CA (2007)

Elliott, M.N., Finch, B.K., Klein, D.J., Ma, S., Do, P., Beckett, M.K., Orr, N., Lurie, N.: Sample designs for measuring the health of small racial ethnic subgroups. Stat. Med. **27**(20), 4016–4029 (2008a). doi:10.1002/sim.3244

Elliott, M.N., Fremont, A.M., Morrison, P.A., Pantoja, P., Lurie, N.: A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. Health Serv. Res. **43**(5p1), 1722–1736 (2008b)

Elliott, M.N., Haviland, A.: Use of a web-based convenience sample to supplement and improve the accuracy of a probability sample. Surv. Methodol. **33**(2), 211–215 (2007)

Falkenstein, M.R.: The Asian and Pacific Islander surname list: as developed from Census 2000. In: Joint Statistical Meetings, New York, NY (2002)

Fiscella, K., Fremont, A.M.: Use of geocoding and surname analysis to estimate race and ethnicity. Health Serv. Res. **41**(4 Pt 1), 1482–1500 (2006)

Fremont, A.M., Bierman, A.S., Wickstrom, S.L., Bird, C.E., Shah, M.M., Escarce, J.J., Rector, T.S.: Use of indirect measures of race/ethnicity and socioeconomic status in managed care settings to identify disparities in cardiovascular and diabetes care quality. Health Aff. **24**(2), 516–526 (2005). doi:10.1377/hlthaff.24.2.516

Fremont, A.M., Lurie, N.: The Role of Race and Ethnic Data Collection in Eliminating Health Disparities. National Academies Press, Washington, DC (2004)

Ghosh-Dastidar, B., Elliott, M.N., Haviland, A., Karoly, L.: Composite estimates from incomplete and complete frames for minimum-MSE estimation in a rare population: an application for families with young children. Public Opin. Q. (in press)

Hand, D.J., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification. Mach. Learn. **45**(2), 171–186 (2001). doi:10.1023/A:1010920819831

Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1), 29–36 (1982)

Institute of Medicine: Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. National Academies Press, Washington, DC (2002)
Jirousek, R., Preucil, S.: On the effective implementation of the iterative proportional fitting procedure. Comput. Stat. Data Anal. **19**(2), 177–189 (1995). doi:10.1016/0167-9473(93)E0055-9
Kestenbaum, B.B., Ferguson, R., Elo, I., Turra, C.: Hispanic identification. In: Southern Demographic Association Meetings, New Orleans, LA (2000)
Lauderdale, D., Kestenbaum, B.B.: Asian American ethnic identification by surname. Popul. Dev. Rev. **19**(3), 283–300 (2000)
Logan, J.: Ethnic Diversity Grows, Neighborhood Integration Lags Behind. Lewis Mumford Center, University at Albany, Albany, NY (2001)
Massey, D.S., Denton, N.A.: Hypersegregation in U.S. metropolitan areas: black and hispanic segregation along five dimensions. Demography **26**(3), 373–391 (1989). doi:10.2307/2061599
McCaffrey, D., Elliott, M.N.: Power of tests for a dichotomous independent variable measured with error. Health Serv. Res. **43**(3), 1085–1101 (2008). doi:10.1111/j.1475-6773.2007.00810.x
Morrison, P.A., Word, D.L., Coleman, C.D.: Using first names to estimate racial proportions in populations. In: Population Association of America Annual Meeting, Washington, DC (2001)
National Health Plan Collaborative: Phase 1 summary report: reducing racial and ethnic disparities improving quality of health care. Hamilton, NJ (2006)
National Research Council: Eliminating Health Disparities: Measurement and Data Needs. National Academies Press, Washington, DC (2004)
Perkins, R.C.: Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results. U.S. Census Bureau, Population Division (1993)
Schenker, N., Parker, J.D.: From single-race reporting to multiple-race reporting: using imputation methods to bridge the transition. Stat. Med. **22**(9), 1571–1587 (2003). doi:10.1002/sim.1512
U.S. Office of Management of Budget: Revisions to the standards for the classifications of federal data on race and ethnicity. Notice. Federal Register, Washington, DC (1997)
Word, D.L., Coleman, C.D., Nunziata, R., Kominski, R.: Demographic aspects of surnames from Census 2000. Available at: http://www.census.gov/genealogy/www/surnames.pdf (2008). Accessed 30 July 2008