DOI: 10.1111/j.1475-6773.2008.00854.x
RESEARCH ARTICLE

A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity

Marc N. Elliott, Allen Fremont, Peter A. Morrison, Philip Pantoja, and Nicole Lurie

Objective. To efficiently estimate race/ethnicity using administrative records to facilitate health care organizations' efforts to address disparities when self-reported race/ethnicity data are unavailable.

Data Source. Surname, geocoded residential address, and self-reported race/ethnicity from 1,973,362 enrollees of a national health plan.

Study Design. We compare the accuracy of a Bayesian approach to combining surname and geocoded information to estimate race/ethnicity to two other indirect methods: a non-Bayesian method that combines surname and geocoded information and geocoded information alone. We assess accuracy with respect to estimating (1) individual race/ethnicity and (2) overall racial/ethnic prevalence in a population.

Principal Findings. The Bayesian approach was 74 percent more efficient than geocoding alone in estimating individual race/ethnicity and 56 percent more efficient in estimating the prevalence of racial/ethnic groups, outperforming the non-Bayesian hybrid on both measures. The non-Bayesian hybrid was more efficient than geocoding alone in estimating individual race/ethnicity but less efficient with respect to prevalence ($p < .05$ for all differences).

Conclusions. The Bayesian Surname and Geocoding (BSG) method presented here efficiently integrates administrative data, substantially improving upon what is possible with a single source or from other hybrid methods; it offers a powerful tool that can help health care organizations address disparities until self-reported race/ethnicity data are available.

Key Words. Bayes's theorem, health disparities, health plans, race, surname

Efforts to measure, monitor, and address racial/ethnic disparities in health care have been limited by the paucity of data regarding the race/ethnicity of users of the health care system. Indeed, until recently, many viewed the

collection of such data as illegal (Fremont and Lurie 2004). One result is that the preponderance of studies on racial/ethnic differences in quality of care and patient outcomes has been limited to patients enrolled in Medicare or Medicaid. Several reports from the Institute of Medicine and the National Academy of Sciences recommend universal collection of self-reported data regarding race, ethnicity, and socioeconomic status as a first step toward addressing disparities (Institute of Medicine 2002; National Research Council 2004). While self-reported data are widely considered to be the gold standard, absent a mandate to do so, collection of such data will be slow and inconsistent.

Several efforts to collect and use such data are underway. For example, the Health Research and Educational Trust, an independent research affiliate of the American Hospital Association, has developed a toolkit for and is assisting a growing number of hospitals with collection of racial, ethnic, and language data. Similarly, a group of hospitals funded by the Robert Wood Johnson Foundation to address disparities in cardiovascular care have committed to collecting race/ethnicity data and monitoring quality of care for different racial/ethnic groups. State policy has also moved toward collecting racial/ethnic data. For example, as part of the Massachusetts health care reform legislation, collection of race/ethnicity data from all hospitalized patients is required by law (Boston Public Health Commission 2006). In California, SB 853 and related regulations require HMO plans to collect race, ethnicity, and language information (California State Senate 2007). Finally, several of the plans participating in the National Health Plan Collaborative to Improve Quality and Eliminate Disparities have begun voluntary collection of self-reported data on the race/ethnicity of their enrollees (National Health Plan Collaborative 2006). Aetna has the most experience in doing so, but even with a mandate from their CEO and significant investment of resources over the past 4 years, the plan has been able to obtain data on only one-third of their enrollees thus far. Although a few smaller regional plans that followed Aetna's lead have obtained a similar proportion of self-reported data in less time, completing the process will likely take several more years.

Address correspondence to Marc N. Elliott, Ph.D., RAND Corporation, 1776 Main Street, Santa Monica, CA 90407; e-mail: elliott@rand.org. Allen Fremont, M.D., Ph.D., and Philip Pantoja, M.A., are with the RAND Corporation, Santa Monica, CA. Peter A. Morrison, Ph.D., is with the RAND Corporation, Nantucket, MA. Nicole Lurie, MD, is with the RAND Corporation, Arlington, VA.

Surname and Geocoding Approaches

Because the process of obtaining self-reported race/ethnicity data can take years to complete, investigators have developed methods of estimating race/ethnicity indirectly from other sources. Two such methods are geocoding and surname analysis. Geocoding uses an individual's address to link individuals to census data about the geographic areas where they live. For example, knowing that a person lives in a Census Block Group (a small neighborhood of approximately 1,000 residents) where 90 percent of the residents are African American provides useful information for estimating that person's race.

Surname analysis infers race/ethnicity from surnames (last names). Insofar as a particular surname belongs almost exclusively to a particular group (as defined by race, ethnicity, or national origin), it is possible to identify its holder's probable membership in the group by using well-formulated surname dictionaries. Such dictionaries now exist for identifying Hispanics and various Asian nationalities (Perkins 1993; Abrahamse, Morrison, and Bolton 1994; Kestenbaum et al. 2000; Lauderdale and Kestenbaum 2000; Falkenstein 2002). Separate surname lists have been generated for Chinese, Indian, Japanese, Korean, Filipino, and Vietnamese Americans. Experimental dictionaries for identifying Arab Americans are under development (Morrison et al. 2003). Both surname analysis and geocoding have recognized limitations—the former has almost no ability to distinguish blacks from non-Hispanic whites whereas the latter has little ability to identify Hispanics or Asians. Although these limitations have been partially overcome by combining the two approaches, the accuracy of prior combined approaches varies widely by geographic area, depending on the prevalence and degree of segregation of racial/ethnic groups (Fremont et al. 2005; Fiscella and Fremont 2006).

A New Hybrid Approach

To further address limitations of current indirect estimation approaches, we developed a new hybrid approach using Bayes's theorem. Bayes's theorem is commonly applied to medical diagnostic testing; in the context of evaluating diagnostic tests, the probability of a given individual having a disease depends both upon (1) an individual's prior probability of having the disease (usually determined from a base rate appropriate to the individual's risk group) and (2) the result of a diagnostic test. Bayes's Theorem updates prior probabilities with test results by considering the *sensitivity, Se* (probability of a positive test result

for a positive individual), and *specificity*, *Sp* (probability of a negative test result for a negative individual), of the diagnostic test to produce an updated (posterior) probability, called the *positive predictive value, PPV*, that efficiently incorporates both sources of information using the formula:

$$PPV = P \times Se/(P \times Se + (1 - P) \times (1 - Sp))$$

Here, we extend the approach from the two-category prior probability that characterizes baseline disease prevalence rates and treat the racial/ethnic distribution of where an individual lives as a four-category prior, the categories being Hispanic, African American, Asian, and non-Hispanic white or other. Our "baseline prevalence" is based on the racial/ethnic composition of the Census Block Group to which the residence of the individual was geocoded. We treat the combined results of the Census Bureau Spanish Surname List and the Lauderdale–Kestenbaum Asian Surname List as another diagnostic test with three possible outcomes (surname appears on Asian list regardless of appearance on Hispanic list, surname appears on Spanish but not Asian list, surname appears on neither surname list).

Using a more general form of Bayes's Theorem, we then use the surname lists to update the prior probabilities of membership in each of the four race/ethnic categories with the surname list results to produce efficient, updated posterior probabilities of membership in the four groups. The extent of this updating increases with the sensitivity and specificity of the surname lists for the population in question. We refer to this new hybrid method as the *Bayesian Surname and Geocoding* method (BSG) to note that it uses a Bayesian approach to combine surname and geocoded information. These probabilities, in turn, can be used to estimate racial/ethnic composition. Though not the focus of the current validation analyses reported here, the estimates can also be used to identify possible disparities in health care or in health outcomes by race/ethnicity.

We compare the accuracy of BSG in estimating race/ethnicity to two other approaches, in all instances evaluating performance against a gold standard of self-report. The first alternative approach is a previous algorithm for combining the two information sources (Fremont et al. 2005; Fiscella and Fremont 2006) that we will here call the *Categorical Surname and Geocoding* approach (CSG) in order to note that it combines surname and geocoded information in a categorical fashion, described below. The second approach to which we compare BSG is one based solely on the geocoded racial/ethnic composition of the Census Block Group where each member lives. We call

Table 1: Summary of Three Methods Compared

Method	*Needs/Uses Surnames*	*Needs/Uses Addresses*	*How It Works*	*Output*
BSG	Yes	Yes	Uses surname lists to update geocoded information and derive posterior probabilities	Probability
GO	No	Yes	Uses geocoded probabilities directly	Probability
CSG	Yes	Yes	Classifies Asians and Hispanics using surname lists; classifies others according to prevalence of blacks in block group	Classification

this final strategy the *Geocoding Only* (GO) approach. These three approaches are summarized in Table 1.

METHODS

Data

We used national enrollment data from Aetna, a large national health plan. The data set consists of self-reported race/ethnicity (as a "gold-standard" used for validation), surname, geocoded address of residence (Census 2000 Block Group level, using the SF1 file), and gender for all 1,973,362 enrollees who voluntarily provided this information to the plan for quality monitoring and improvement purposes. While voluntarily reported race/ethnicity was predominantly non-Hispanic white or other (78.1 percent), the data set included a reasonable distribution of Hispanics (8.9 percent), blacks (8.0 percent), and Asians (5.0 percent); 51.2 percent (1,010,043) were female. Data disclosed to RAND were done so in compliance with HIPAA regulations.

Implementation of the BSG

The Appendix S1 describes the implementation of the BSG algorithm in detail. If the BSG produced classifications instead of probabilities, we could describe its performance in terms of the sensitivity and specificity of the BSG. Instead, we use alternative measures described below. The sensitivities and specificities of the *surname lists* do play a role with BSG, however. They are *inputs* or tuning parameters that determine how the geocoded and surname data are combined to produce posterior probabilities, as detailed in the Appendix S1 (the greater the sensitivity and specificity, the more the surname results change the probabilities derived from geocoding). Thus these surname

list sensitivities and specificities are not directly evaluative of performance in this context, but are primarily intermediate parameters.

As applied to the primary data set, the sensitivity of the Spanish and Asian surname lists themselves were calculated at 80.4 and 51.5 percent, respectively. The specificities are 97.8 and 99.6 percent, respectively. These sensitivities and specificities are characteristics of the surname lists, not of the BSG. Table S1 describes the probability of members of a given group appearing on each surname list or neither given these sensitivities and specificities. For example, Asians will appear on the Asian list 51.5 percent of the time (irrespective of appearance on the Spanish list), on the Spanish list but not the Asian list 1.1 percent of the time, and on neither list 47.4 percent of the time at these levels of sensitivity and specificity under the assumptions stated earlier.

Because we find higher sensitivity for males than females (83.1 versus 77.8 percent on the Spanish Surname List; 52.7 versus 50.2 percent on the Asian Surname List, $p<.05$ for each) and slightly higher specificity for males than females for the Spanish Surname List (98.0 versus 97.5 percent, $p<.05$) that are presumably related to retention of surnames after marriage, the BSG uses gender-specific sensitivities and specificities. Thus, for example, a male who appears on the Spanish surname list in a given block group receives a slightly higher posterior probability of being Hispanic than a female who appears on that same list from the same block group because the surname list is known to be more accurate for males than females. The Appendix S1 provides additional examples of how the BSG generates posterior probabilities as well as other details of its implementation.

Other Algorithms Used for Comparison with the BSG

The second method, GO, simply uses the racial/ethnic prevalences from Census Block Groups as probabilities. Surname lists provide no means by which to distinguish blacks from non-Hispanic whites, so do not permit estimates of disparities between these two groups. For this reason, a "surname only" approach is not considered.

Instead, we consider a previously described alternative combination of geocoding and surname information, the CSG (Fiscella and Fremont 2006). CSG categorizes individuals through a series of steps. It (1) labels a person Hispanic if their name appears on the Spanish surname list; if not, it (2) labels a person Asian if the name appears on the Asian surname list; if neither of these applies, geocoded race/ethnic information is used to adjudicate classifications

among the remaining individuals into black or non-Hispanic white categories. In particular, (3) if an individual not appearing on either surname list resides in a block group that is at least 66 percent black, they are classified as black; (4) otherwise they are classified as non-Hispanic white. In an application using Medicare enrollees in a national health plan, this algorithm produced estimates of racial/ethnic heath disparities that were similar to those obtained with self-reported race-ethnicity (Fremont et al. 2005; Fiscella and Fremont 2006).

Outputs of BSG, CSG, and GO: Classifications versus Probabilities

CSG discretely classifies each plan member into one of four racial/ethnic categories, whereas BSG and GO produce probabilities of membership in each of these four groups. As an illustration, consider a hypothetical Bob Jones living in a Census Block Group that was 67 percent white/other, 11 percent black, 11 percent Hispanic, and 11 percent Asian. CSG would note that "Jones" was on neither surname list and that his block group was <66 percent black and would therefore classify Mr. Jones as white/other. GO would simply use these four prevalences as probabilities and estimate that Mr. Jones had a 67 percent chance of being white/other and an 11 percent chance of being a member of each of the other three groups. As illustrated in Table 2, BSG would note that "Jones" was on neither surname list and integrate that information with the sensitivities and specificities of those lists, as well as the racial/ethnic composition of his block group to estimate that Mr. Jones has a 78.7 percent chance of being white/other, a 12.9 percent chance of being black, a 6.1 percent chance of being Asian, and a 2.2 percent chance of being Hispanic. Note that being on neither surname list makes white/other and black more likely than they were before surnames were considered, and that the probability of being Hispanic falls more than the probability of being

Table 2: Illustration of BSG Posterior Probabilities of the Race/Ethnicity of a Male Individual Living in a Census Block Group That Was 67 Percent White/Other and 11 Percent Each Asian, Hispanic, and Black

	BSG Posterior Probability of Race/Ethnicity			
Surname	*Asian*	*Hispanic*	*Black*	*White/Other*
Wang	0.937	0.008	0.008	0.048
Martinez	0.010	0.845	0.021	0.125
Jones	0.061	0.022	0.129	0.787

Asian (because the Spanish surname list has greater sensitivity than the Asian list). Additional examples appear in Table S3.

One can estimate prevalences, means, and disparities by race/ethnicity by working directly with probabilities, without ever producing individual classifications. For example, if one's goal were a prevalence estimate, averaging probabilities is more accurate than classifying and rounding before summing (McCaffrey and Elliott forthcoming). For example, in an area with 10 people who had a 57 percent chance of being white and a 43 percent chance of being black and another 10 people with a 69 percent chance of being white and a 31 percent chance of being black, racial/ethnic prevalences would be more accurately estimated as 63 percent white and 37 percent black (averaging probabilities) than as 100 percent white (classifying each person into the group that was most likely for them). Please see Table S4 for additional examples. Similarly, if the goal is to compare racial/ethnic groups in terms of a clinical process measure, such as adherence to diabetes care recommendations as measured by administrative records, one need not classify individuals into discrete categories. Instead, one can enter an individual's probabilities of membership in each of several racial/ethnic groups (omitting one as a reference group) as predictors in a linear or logistic regression and the coefficients will be unbiased estimates of the difference of each racial/ethnic group from the reference racial/ethnic group in the outcome. Moreover, McCaffrey and Elliott show that such direct use of these probabilities, while less accurate than truly knowing race/ethnicity with certainty for each individual, is more accurate and efficient than using categorical classifications based on these probabilities. In each of these instances, categorizing continuous probabilities into discrete classifications is an unnecessary step that discards substantial information by ignoring distinctions in probabilities. While there may be some instances in which one must make a discrete decision for specific individuals (e.g., whether to mail Spanish-language materials to specific addresses), direct use of probabilities will be more efficient for aggregate statistical inferences, including the comparison of racial/ethnic groups.

If we were only examining CSG, we could describe its accuracy of classification in terms of sensitivity, specificity, and positive predictive value. Because we are comparing both classification-based and probability-based methods, we employ different performance measures.

Evaluation

We compare BSG, CSG, and GO in terms of how closely the estimates of race/ethnicity that they produce match those derived from self-reported race/

ethnicity for the same individuals. We develop two performance metrics applicable to all three approaches (BSG, CSG, and GO). We then compare the relative efficiency of the three methods according to these two metrics. The first metric assesses accuracy in matching the four-category distribution of self-reported racial/ethnic prevalence in a population. The second metric assesses the accuracy of predicting individual race/ethnicity—the extent to which those who self-report a given race/ethnicity are assigned higher probabilities of that race/ethnicity (or are more likely to be classified as that race/ethnicity). The two measures are complementary in that the first detects systematic errors in four-category classifications (e.g., a method is overly likely to classify someone as white and insufficiently likely to classify someone as black), and the second measure detects unsystematic errors (e.g., a method doesn't overestimate or underestimate any group in aggregate, but is just not very accurate in predicting the race/ethnicity of specific individuals).

Performance Metric for Predicting Racial/Ethnic Prevalence

For each of the three methods, we report the prevalence estimates derived for each of four racial/ethnic groups and compare these with self-reported proportions. In order to summarize the accuracy across these four categories, we compute the average error of the four categorical racial/ethnic prevalences estimates, weighted by their true (self-reported proportions). Ratios of average squared errors can be used to measure the *relative efficiency* of two methods in estimating prevalences. To say that method one has a relative efficiency of 3.0 relative to method two means that the accuracy of method one using a given sample size is the same as what would be obtained with three times the sample size using method two.

Performance Metric for Predicting Individuals' Race/Ethnicity

The Brier score (Brier 1950) is the mean squared deviation of a prediction from the true corresponding dichotomous outcome. The Murphy decomposition of the Brier score (Yates 1982) distinguishes (a) uncontrollable variation due to the prevalence of the outcome from (b) the extent to which predictions correlate with the dichotomous outcome. We use this correlation (b) as our measure of performance in predicting individual race/ethnicity. This metric rescales predictive performance to a $(0, 1)$ scale regardless of prevalence.

In particular, we use the correlation of the dichotomous or probabilistic prediction with a dichotomous indicator of true self-reported race-ethnicity for each of four racial/ethnic groups. Whether a method produces classifications

or probabilities, it is a comparable measure of the accuracy with which individual race/ethnicity is predicted. Estimates for the four racial/ethnic measures are not independent, but are negatively correlated. To summarize performance across all four racial/ethnic categories, we also calculate an average correlation, weighted by prevalence, for each method. By comparing ratios of squared correlations, we can compare the relative efficiency of methods in predicting individual race/ethnicity.

RESULTS

Predicting Racial/Ethnic Prevalences: Comparing BSG, CSG, and GO

Table 3 displays the overall proportions of self-reported race/ethnic data falling into the four categories, along with estimates derived from each of the three methods using the primary data set. The average deviation from self-report is also displayed for each method. When comparing methods, it may be noted that the sampling error in assessing accuracy in prevalence is sufficiently small that all differences of 0.1 percent or more are statistically significant. GO substantially overestimates the prevalence of Hispanics, moderately overestimates the prevalence of blacks, and moderately underestimates the prevalence of Asians ($p < .05$ for each). CSG is very accurate for Hispanics, but it underestimates the prevalence of Asians by nearly a factor of two and underestimates the prevalence of blacks by nearly a factor of three ($p < .05$ for both). These patterns result in overestimating the proportion of plan members who are white.

BSG is the most accurate overall, with a weighted average prevalence error (deviation from self-reported) of 1.6 percent, followed by 2.0 percent for

Table 3: Comparing Overall Racial/Ethnic Prevalence Estimates to Self-Report Estimates ($n = 1{,}973{,}362$)

	Estimated Percentage in Each Group				*Weighted Average Overall Deviation from Self-Report*
	Hispanic	*Asian*	*Black*	*White/Other*	
SELF-REPORT	**8.9**	**5.0**	**8.0**	**78.1**	(0)
BSG	10.0	4.5	9.1	76.4	1.6%
GO	10.8	4.2	9.0	76.0	2.0%
CSG	9.2	2.9	3.0	84.9	6.2%

Ninety-five percent margins of sampling error are <0.1% for a single prevalence estimate, a difference in prevalences estimates across methods.

GO and 6.2 percent for CSG ($p<.05$ for all pairwise comparisons). BSG moderately overestimates Hispanic and black prevalence, while underestimating whites and Asians somewhat ($p<.05$ for each). BSG is 56 percent more efficient than geocoding alone in prevalence estimates, whereas CSG is less efficient for this purpose than geocoding alone.

Predicting Individual Race/Ethnicity: Comparing BSG, CSG, and GO

Table 4 displays the correlation with self-reported race/ethnicity for each of the three methods and four race/ethnic groups in the primary data set. All reported correlations are statistically significant and differ across methods at $p<.05$. BSG predictions correlate with individual indicators of race/ethnicity at 0.61 to 0.79, with a weighted average correlation of 0.70.

CSG is the next best by this measure (average correlation 0.63), with similar performance for Hispanics and Asians, somewhat lower performance for whites, and notably lower performance for blacks. GO (average correlation 0.53) was near the performance of the BSG and notably better than CSG for blacks, but performed less well than the other two algorithms for all other groups, performing especially poorly for Hispanics and Asians. Overall, BSG was 74 percent more efficient than geocoding alone in estimating individual race/ethnicity and CSG was 41 percent more efficient than geocoding alone in predicting individual race/ethnicity. This means that 1,000 observations from BSG provide as much information as 1,740 observations using geocoding alone. For Hispanics and Asians, BSG has 2.6 and 3.9 times the efficiency of geocoding alone, respectively.

BSG performed better than each of the alternatives by both performance metrics and increases efficiency by 56–74 percent relative to geocoding alone. In contrast, the CSG improves upon direct use of geocoded data by only one

Table 4: Correlation of Individual Predicted Race/Ethnicity with Self-Reported Race/Ethnicity ($n = 1{,}973{,}362$)

	Correlation with Self-Reported Race/Ethnicity				
	Hispanic	*Asian*	*Black*	*White/Other*	*Weighted Average*
BSG	0.79	0.67	0.61	0.70	0.70
GO	0.49	0.34	0.57	0.55	0.53
CSG	0.77	0.65	0.48	0.63	0.63

All differences in correlations by methods are significant at $p<.05$.

of these metrics, highlighting the importance of *how* surname and geocoded information is combined.

DISCUSSION

We have described a method for estimating race/ethnicity using administrative data. This approach, which applies Bayes's Theorem to a four-category geocoding and surname analysis, appears to be a particularly useful means of integrating these sources of information and substantially outperforms a classification-based means of combining this information (CSG). The advantage of BSG over CSG probably stems from two factors: (1) better identification of blacks in areas of low residential segregation and (2) greater precision through the direct use of probabilities.

In addition to its ability to estimate race/ethnicity, the BSG approach has substantial potential for use in routine assessment and monitoring of health disparities in a population. It can also be used when estimated race/ethnicity is to be a predictor in multivariate regression or other models; thus its usefulness is not limited to estimation of disparities or to health applications.

One limitation, which applies to all methods of inferring race/ethnicity, is that while BSG supports modeling at the individual level, it is not accurate enough to support individual-level interventions and requires large sample sizes for good precision, because there is some inherent loss of information compared with self-reported race/ethnicity for a sample of the same size. Secondly, although results were evaluated on a large, racially and ethnically diverse national sample, results may differ somewhat for those not insured by this health plan or those who do not self-report race-ethnicity.

An additional limitation is that the direct use of predicted probabilities is somewhat more complex than the use of 1/0 categorical indicators of race/ethnicity and may be unfamiliar to some analysts. Traditionally, analysts have either used a single categorical variable with each level representing a particular racial/ethnic group, or a series of "dummies," that is—separate variables (one for each race/ethnicity) that have a value of "0" if the person is not, for example, Asian, or "1" if the person is Asian. The posterior probabilities from the BSG and GO are continuous variables with values from 0 to 1 that are used somewhat differently. Nonetheless, this approach is still relatively straightforward, and one can interpret the coefficients as if they were from racial/ethnic dummy variables. The Appendix S1 provides examples of how these probabilities can be used within *SAS*.

Our new method of estimating race/ethnicity substantially outperforms other widely used indirect methods and provides health plans and others a timely means to infer race/ethnicity among plan members. Although self-reported race/ethnicity represents a gold standard in many situations, indirect methods offer a powerful and immediate alternative for estimating health experiences by racial/ethnic status using only administrative data. In combination with geographic information systems (GIS) tools, these methods can be of great use to health plans, researchers, and others (National Health Plan Collaborative 2006).

Future work can directly examine the accuracy of BSG in estimating health disparities, as well as seek further improvements in the accuracy of BSG estimates of race/ethnicity. One way to do the latter might be to develop regional sensitivity and specificity parameters. Such data would also provide insight into the extent to which BSG performance varies by plan or region. One could model racial-ethnic selection into health insurance within Block Group conditional on surname results, further improving BSG performance (because our results imply there are lower rates of health coverage for blacks and Hispanics than for Asians and whites/others even within the same Block Groups).

Finally, when applying BSG to a specific population, such as a commercially insured population, one could use Census racial/ethnic data within block groups that were restricted to ages that better matched the target population. To the extent that age differed by race/ethnicity, this would further reduce BSG bias and improve its performance. Future work should follow along these paths to refine an already promising and useful approach to inferring race/ethnicity from names and addresses alone.

ACKNOWLEDGMENTS

This study was supported, in part, by contract 282-00-0005, Task Order 13 from DHHS: Agency for Healthcare Research and Quality. Marc Elliott is supported in part by the Centers for Disease Control and Prevention (CDC U48/DP000056). The authors would like to thank Kate Sommers-Dawes and Scott Stephenson for assistance with the preparation of the manuscript.

Disclaimers: The contents of the publication are solely the responsibility of the authors and do not necessarily reflect the official views of the CDC.

Disclosure: None

REFERENCES

Abrahamse, A. F., P. A. Morrison, and N. M. Bolton. 1994. "Surname Analysis for Estimating Local Concentration of Hispanics and Asians." *Population Research and Policy Review* 13: 383–98.

Boston Public Health Commission. 2006. *Data Collection Regulation.* Boston: Boston Public Health Commission.

Brier, G. W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

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

Falkenstein, M. R. 2002. "The Asian and Pacific Islander Surname List: As Developed from Census 2000." Paper read at Joint Statistical Meetings.

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.

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

Fremont, A. M., P. Pantoja, M. N. Elliott, P. A. Morrison, A. F. Abrahamse, and N. Lurie. 2005. "Use of Indirect Measures of Race/Ethnicity to Examine Disparities in Managed Care. AcademyHealth Annual Research Conference. Chicago, IL.

Institute of Medicine. 2002. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.* Washington, DC: The National Academies.

Kestenbaum, B. B., R. Ferguson, I. Elo, and C. Turra. 2000. "Hispanic Identification." Paper read at 2000 Southern Demographic Association Meetings.

Lauderdale, D., and B. B. Kestenbaum. 2000. "Asian American Ethnic Identification by Surname." *Population and Development Review* 19 (3): 283–300.

McCaffrey, D., and M. N. Elliott. 2007. "Power of Tests for a Dichotomous Independent Variable Measured with Error." *Health Services Research.* DOI: 10.1111/j.1475-6773.2007.00810.x

Morrison, P. A., B. Kestenbaum, D. S. Lauderdale, A. F. Abrahamse, and S. El-Badry. 2003. "Developing an Arab-American Surname List: Potential Demographic and Health Research Applications." Paper read at 2003 Southern Demographic Association Meetings.

National Health Plan Collaborative. 2006. *Phase 1 Summary Report: Reducing Racial and Ethnic Disparities and Improving Quality of Health Care.* Hamilton, NJ: National Health Plan Collaborative.

National Research Council. 2004. *Eliminating Health Disparities: Measurement and Data Needs.* Washington, DC: National Academies Press.

Perkins, R. C. 1993. "Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results.": U.S. Census Bureau, Population Division.

Yates, J. F. 1982. "External Correspondence: Decompositions of the Mean Probability Score." *Organizational Behavior and Human Performance* 30: 132–56.

SUPPLEMENTARY MATERIAL

The following material is available for this article online:

Appendix S1: Implementation of Bayesian Surname and Geocoding Combination (BSG).

Appendix S2: Author Matrix.

Table S1: Probabilities of Joint Surname Test Results by True Race/Ethnicity.

Table S2: Posterior Probabilities of Group Membership by Surname List Results.

Table S3: BSG Posterior Probabilities of Race/Ethnicity for Hypothetical Block Groups A, B, C, D, and E for Males ($N = 963,319$).

Table S4: Example of BSG, GO, and CSG estimates of Racial/Ethnicity of Plan Membership in a Hypothetical Block Group (67 Percent White/Other, 11 Percent Each Black, Hispanic, and Asian Overall), Based on 10 Male# Plan Members (2 on Asian Surname List, 3 on Spanish Surname List, 5 Unlisted).

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1475-6773.2008.00854.x (this link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.