A NEW METHOD FOR ESTIMATING RACIAL/ETHNIC DISPARITIES WHERE ADMINISTRATIVE RECORDS LACK SELF-REPORTED RACE/ETHNICITY [DRAFT- THIS IS WORK IN PROGRESS]

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

RAND
3/26/06

ABSTRACT [150-word limit]
With increasing ethnic diversity, equity is a continuing focus of policy formulation and political debate. We consider the need by health plans to monitor racial/ethnic disparities in health care quality among their enrollees. Few plans acquire racial/ethnic data from their entire membership. Where classification variables are missing, individuals' surnames and neighborhood contextual measures can provide useful surrogate data elements for comparing population subgroups. Building on the strengths of surname analysis and neighborhood contextual analysis, we present and evaluate a hybrid method which is broadly applicable where researchers must rely on administrative records lacking racial/ethnic detail. This Bayesian Algorithm integrates both sources of information and substantially outperforms other approaches. It performs well when race/ethnic classification is the only goal or when estimated race/ethnicity is to be a predictor in regression or other models. Thus its potential applications are not limited to estimation of disparities or to health applications.

PRELIMINARY DRAFT PAPER

INTRODUCTION

As the nation and local communities grow more ethnically diverse, issues of equity have become a continuing focus of policy formulation and political debate. Such issues cover a broad spectrum: access to higher education, housing, employment opportunities, and health care to name a few (see, for example, Morrison, 2003a; Clark and Morrison, 2005; Morrison, forthcoming). Addressing such issues necessitates classifying people into various population subgroups defined by race, ethnicity, and national origin. Self-reported data, the gold standard, are often unavailable, and in their absence demographers occasionally are called upon to devise practical ways to distinguish population subgroups for purposes of gauging equity across racial/ethnic groups.

Our proposed paper focuses on one such instance: the need by health plans to monitor racial/ethnic disparities in health care quality among their enrollees. Spurring

the recent momentum behind insurers' and employers' concerns have been federal health officials and physician groups representing both black and Hispanic doctors and several national reports (Institute of Medicine, 2002; National Quality Forum 2002;Workgroup on Quality: National Committee on Vital and Health Statistics 2004). Yet, few health plans systematically gather racial/ethnic data from their entire membership. Where classification variables are missing, individuals' surnames and neighborhood contextual measures can provide useful surrogate data elements for comparing population subgroups.

BACKGROUND

Health plans generally do not retain racial/ethnic data on enrollees because of uncertainty about the legality of collecting such data or fear that consumers would assume plans were using this information inappropriately (Fremont and Lurie 2004). Plans have limited options for quickly obtaining needed data. Information about race and ethnicity could be solicited at the time of enrollment for *new* members, but alternative strategies are needed to obtain this information from *existing* plan members. Potential direct methods include mail, telephone, or internet surveys; onsite collection at point of care; and supply by employers, hospitals, states, or Centers for Medicare and Medicaid Services. However, each of these strategies has limitations in terms of reliability, validity, bias, and completeness. Even using a combination of these direct methods, the process will take most health plans years to obtain race/ethnicity data on the bulk of their enrollees.

As a practical matter, no method for obtaining race and ethnicity data can be entirely accurate or bias free. Self reports may be limited by nonresponse. Furthermore, hospitals' inconsistent collection and classification of race and ethnicity data is potentially problematic. For example, one study reports that race was coded differently upon re-hospitalization for 6% of re-hospitalized African Americans and 11% of re-hospitalized whites (Blustein 1994). Another study found that race and ethnicity compiled by the Veterans Administration Health System corresponds with self-report data only 60% of the time, with lower agreement for non-whites and better educated patients (Kressin et al. 2003).

RAND is supporting a group of national health insurance plans that are examining quality of care for racial/ethnic groups, and pilot testing interventions to reduce noted disparities. The plans needed to estimate the race and ethnicity of their enrollees in order to examine quality of care for various groups. We have drawn on two familiar approaches to meeting these needs: surname analysis and neighborhood contextual analysis (see Fremont, et al., 2005; and forthcoming, for elaboration). Building on the strengths of each approach, we have devised and evaluated a new hybrid method which is broadly applicable where researchers address issues of equity using administrative records lacking racial/ethnic detail.

Available Methods

Surname analysis encompasses techniques for estimating the membership of particular racial and ethnic communities within a population. Insofar as a particular surname belongs almost exclusively to a particular (racial, ethnic, national origin) group,

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 (see Abrahamse, et al., 1994; Falkenstein, 2002; Kestenbaum et al., 2000; Lauderdale and Kestenbaum, 2000; Perkins, 1993). Experimental dictionaries for identifying Arab Americans are under development (Morrison et al, 2003b).[1]

Neighborhood contextual analysis encompasses techniques for estimating individuals' race/ethnicity from local areal measures of group composition. Geo-coding is a common approach marketers use to assign group membership to individuals probabilistically. Geocoding involves using plan members' addresses to identify geographic areas where they live and linking this information to census data about that area. Because sociodemographic characteristics of communities correlate with characteristics of the residents who live there, geocoded measures can be used to infer characteristics about persons living in those areas. For example, knowing that a person resides on a census block where 90% of the residents are African American provides information for estimating that person's race, given certain assumptions about racial residential separation.

Limitations of Each Method

Both surname analysis and neighborhood contextual analysis have recognized limitations. In practice, a surname never is an exact identifier of its bearer's ethnicity. With Spanish surname analysis, for example, not all Spanish-surnamed persons self-identify as Hispanic; conversely, not all self-identified Hispanic persons have Spanish surnames. Relying on a list of names to infer Hispanic ethnicity, then, exposes one to

[1] There are several ways to assign ethnicity based on names, including the use of letter combinations, dictionaries of surnames, and combinations of first, middle, and last names. The original approach, the Generally Useful Ethnicity Search System (GUESS), was developed using 1953 California Department of Public Health birth data (Perez-Stable et al. 1995). The program was derived using an algorithm based on common Spanish names, given name, and mother's maiden name. It uses the linguistic structure of the last name to assign Hispanic ethnicity. GUESS was updated in the 1980s using more current Spanish surnames (Rosenwaike and Bradshaw 1988). A simpler and more commonly used approach is to assign Hispanic ethnicity based on a surname list. Such a list was developed using 1980 Census data(Perkins 1993) and then revised using 1990 data.(Word and Perkins 1996). Surname lists have also been used to identify Asian subpopulations in the United Kingdom (UK) (Harland et al. 1997;Nanchahal et al. 2001;Nicoll, Bassett, and Ulijaszek 1986), Australia (Hage et al. 1990), Canada (Choi et al. 1993;Coldman, Braun, and Gallagher 1988;Sheth et al. 1997) and the US (Lauderdale D.S. and Kestenbaum B 2000;Swallen, Glaser, Stewart, West, Jenkins, and McPhee 1998). The best validated list has been produced by Lauderdale and Kestenbaum using the Social Security Administration's file (Lauderdale D.S. and Kestenbaum B 2000). Separate surname lists have been generated for Chinese, Indian, Japanese, Korean, Filipino, and Vietnamese Americans.

two types of errors: (1) "false positives," e.g., classifying a non-Hispanic person as "Hispanic" because his or her surname happens to be on the Census Bureau List of Spanish Surnames; and (2) "false negatives," i.e., classifying an Hispanic person as "non-Hispanic" because his or her name is not listed. Such misclassifications can arise for various reasons. A woman may relinquish her maiden Hispanic surname or acquire an Hispanic surname from her husband. Alternatively, either or both spouses may elect to hyphenate their surname (e.g., "Lee-Flores" or "Torres-Ohara"), thereby confounding precise classification.

Furthermore, particular surnames can be highly misleading in some neighborhood contexts. Persons with the common surname "Lee," for example, are likely to be Korean or Chinese if they reside in a predominantly Asian neighborhood but not if they live in, say, Williamsburg, Virginia. Likewise, the Asian surname "Ohara" could easily misidentify persons living in predominantly Irish neighborhoods.

Lastly, surname lists do not distinguish one important racial group (blacks) from other racial groups in the population. This is a critical limitation where (as in our application) the aim is to monitor racial/ethnic disparities in health care quality.

Neighborhood contextual analysis also has serious limitations stemming from the well-known "ecological fallacy." Area-level data (e.g., census tract, block group, or block data) are informative only where micro-segregation is strong. Knowing that an area (e.g., census block) is populated equally by each of four different groups is of little value in estimating the identity of a particular inhabitant.[2]

Proposed Hybrid Method

Given the limitations above, we have devised and refined a hybrid method which builds on the strengths and possibilities each method offers. Our effort was driven by practical considerations: We needed to classify the members of health plans into various "minority" categories in order to measure and compare each group's health statuses. This entails distinguishing race (white, black, Asian, etc.) and ethnicity (Hispanic, non-Hispanic). Surname analysis offers a feasible approach to identifying Asian or Hispanic group membership, but not black group membership. Conversely, neighborhood contextual analysis offers an effective approach to identifying black group membership in certain contexts. In particular, this method works best when micro-segregation is strong, and blacks are the group that exhibits the greatest degree of micro-clustering. Accordingly, we pursued the possibility of melding these two different approaches into a more powerful and all-inclusive method for classifying members of health plans into all relevant categories for purposes of studying minority health disparities.

DESCRIPTION OF OUR HYBRID METHOD

The algorithm we devised is intended to provide efficient and unbiased estimates of race-ethnic disparities where patient-reported race/ethnicity is unknown, on the basis

[2] One measure of the informativeness of distributional information here is the sum over i of (p(i)^2), where p(i) are the proportions in each racial/ethnic group, i=1,2,….k. 0 is least information, 1 is most.

of address and surname.

We apply a well-known approach from medical diagnostic testing to the present problem. In the diagnostic context, the probability of a given individual having a disease depends both upon the prior probability of their having the disease (usually determined from a base rate appropriate to the individual's risk group) and the result of a diagnostic test. Bayes' 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 validity, PPV*, that efficiently incorporates both sources of information using the formula:

- PPV=P*Se/(P*Se+(1-P)*(1-Sp))

. In the typical diagnostic situation there are only two categories of individuals (have disease, do not) and one test.

Here, we treat the race/ethnic distribution of where an individual lives as a four-category prior probability, analogous to base rates of disease states. The four categories are Hispanic, African-American, Asian, and non-Hispanic white or other. Our "baseline prevalence" is based on the geo-coded distribution by Census block group of individual residence.

We treat the combined results of the Census Bureau Spanish Surname List and the Lauderdale-Kestenbaum Asian Surname List as a single diagnostic test with three possible outcomes (surname appears on Asian list regardless of Hispanic result, surname appears on Spanish but not Asian list, surname appears on neither surname list). Using a more general form of Bayes' Theorem, we may update the prior probabilities of membership in each of the four race/ethnic categories with the results using these surname lists to produce efficient, updated posterior probabilities of membership in the four groups.

These probabilities, in turn, can be used directly in regression analyses to estimate race/ethnic disparities from address and name information alone. We then compare the accuracy of classification and estimates of disparities to geo-coded information alone and to less efficient combinations of the two information sources. The Appendix describes the implementation of the algorithm in detail.

Data

Results thus far are based primarily on analyses of 1,821 enrollees from one major health plan, for whom we have self-reported race/ethnicity (for validation), surname and geocoded address of residence (Census 2000 block group level). Data also include six dichotomous HEDIS indicators of health plan performance. Self-reported race/ethnicity was predominantly non-Hispanic White or Other (68.2%), but also included reasonable representation of Hispanic (11.1%), Asian (10.5%), and Black (10.2%)

Our final paper will report results for a considerably larger sample, drawn from several health plans. Analyses currently in progress, but not reported fully here, include 9,991 observations from two major health plans. This second set, of which the first set is a subset, has a population that is considerably more Hispanic, somewhat less Asian, and somewhat less non-Hispanic White. Except where noted, analyses were performed on

the set of 1,821 cases from the single plan.

We examined the complete set of six dichotomous HEDIS performance measures related to the quality of care for diabetes. Four measures focused on whether specific processes of care were performed: an annual check of HgbA1c, LDL, protein in the urine (an indication of nephropathy), and an eye exam by an ophthalmologist. Two outcome measures were whether HgbA1c and LDL were adequately controlled during the past year. An HgbA1c level at or below 9.5 percent and an LDL level at or below 130 units (mg/dl) indicates control in HEDIS measure specifications.

Evaluation:

Performance metrics

If our only goal were classification, we would need to make a categorical classification of each plan member into one of four racial/ethnic categories. In that case, the accuracy of classification could be described in terms of sensitivity (the percentage of persons of a particular ethnicity in a given population who are correctly coded), specificity (percentage of persons who are not of a particular ethnicity who are correctly coded), and positive predictive value (percentage of persons with a given classification who self-report the ethnicity assigned by the coding method).

Our ultimate goal is the estimation of disparities. For this purpose, it is not necessary to have discrete, categorical classifications. Estimated probabilities of membership in each of the four racial/ethnic groups for each individual are sufficient and can be used directly as predictors in regressions of health outcomes. In fact, it can be demonstrated that discretization of these probabilities (converting them to categorical classifications) involves a loss of information that decreases the accuracy of disparity estimates.

Nonetheless, we are still interested in measures of classification performance because they may identify strengths and shortcoming of the algorithm that affect its performance on the measures of greatest interest. We propose two metrics of classification performance that are general enough to apply to both continuous measures (probabilities) and discrete measures (classifications).

Classification metrics

The first measure of classification performance is the correlation of the classification or probability with a dichotomous indicator of true self-reported race-ethnicity for each of four racial/ethnic groups. In the case of classifications, this is a phi coefficient; in the case of continuous measures, this is a point-biserial correlation. In both cases, it is a comparable measure of the extent to which those of a given race/ethnicity tend to be coded as more likely to be members of that race/ethnicity compared to those not of that race/ethnicity, where 0 represents chance performance and 1 represents perfect performance. Estimates for the four racial/ethnic measures are not independent, but are negatively correlated.

The second measure of classification performance is the accuracy of the estimate of the total proportion of the population that belongs to a given race/ethnicity. In particular, we examine the Mean Squared Error (MSE) of the estimates-the squared difference of the algorithm-estimated proportion of plan members who belong to a given racial/ethnic group from the true proportion of health plan members that belong to a

given racial/ethnic group by self-report. Again, this measure exists for each of the four racial/ethnic groups. The algorithm-estimated proportions are defined as the proportion of plan members classified as belonging to a given racial/ethnic group for discrete methods and the mean probability of belonging to a given racial/ethnic group for continuous measures. This metric is designed to detect biased classification- systematic tendencies to overestimate or underestimate the proportion of a given sample that belongs to a given racial/ethnic group. This can also be thought of as an imbalance of positive and negative predictive validity for a given group.

Metric for disparity estimates

Because our ultimate goal is the estimation of disparities, our most important metric is accuracy in the estimation of these disparities. For each health outcome, we estimate three disparities: comparisons of blacks, Hispanics, and Asians to a reference group of non-Hispanic whites. If all estimates were unbiased, the standard reporting metric would be the standard error of estimate. Because that is not the case, we will examine the mean squared error (MSE) of estimates and the square root of this term, the Root MSE (RMSE). The MSE of these disparity estimates is defined as the expected squared deviation of the value estimated with race/ethnicity derived from the algorithm from the value derived from self-reported race/ethnicity in the same sample. It can be decomposed into the sum of the squared standard error of estimate and the squared bias. The first term reflects unsystematic error and the second term reflects systematic error largely attributable to biased classification.

Other algorithms used for comparison

In order to provide benchmarks to which we might compare the performance of the new *Bayesian Algorithm*, we define two other algorithms. The first alternative, which we designate the *Geocoding Algorithm*, simply uses the racial/ethnic prevalences from the zip codes as probabilities. In other words, it is the first step of the Bayesian Algorithm, but makes no use of the surname lists. If all racial/ethnic groups were equally likely to belong to a health plan after accounting for zip codes of residence, this method would result in unbiased classification (and unbiased estimation of disparities). In practice, this is not likely to hold, and bias will be related to the strength of selection into the health plan by race/ethnicity within zip codes. Note that this limitation applies to the Bayesian Algorithm as well. If one knew the overall race/ethnic proportions within a plan (but not individual race/ethnicity), one could use this to adjust for this selection, but we considered this circumstance sufficiently unlikely that we do not consider it here in greater detail.

If one were to use surname lists alone, there would be no ability to distinguish between blacks and non-Hispanic whites and thus no ability to estimate disparities between these two groups. For this reason, such an approach is not presented. Instead, we present a reasonable alternative combination of geocoding and surname information, which we designated *Sequential Classification*. Sequential classification (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 will be 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% black, they are classified as black; (4) otherwise they are classified as non-Hispanic White.

Bias in "Prevalence Estimates", or Prior Race/Ethnic Probabilities

If all race/ethnic groups were equally likely to belong to a health plan after accounting for block group of residence, Census-based race/ethnic distributions would be unbiased prior probabilities for the race/ethnicity of the health plan member. In practice, this is not likely to hold, as there is likely to be selection by race/ethnicity into health plan membership (or health coverage in general), even within block groups. This will translate into bias, one source of inaccuracy, in methods that rely on these prior probabilities. Here, the Bayesian Algorithm and the Geocoding Algorithm will be particularly affected. In theory, if one knew the overall race/ethnic distributions within a plan, but not individual race/ethnicity, one could use this to adjust for overall selection, but such a circumstance seems unlikely.

RESULTS/DISCUSSION

Sensitivity and Specificity of the Joint Surname Test (within the Bayesian Algorithm)

Based on all 9,991 observations currently available, the sensitivity of the Spanish and Asian surname lists are 77.5% and 47.0% respectively. The corresponding specificities are 97.3% and 99.5%. Table 1 below demonstrates the probability of the three joint surname test outcomes under these sensitivities and specificities. As can be seen, Asians will appear on the Asian list 47% of the time (irrespective of appearance on the Spanish list), on the Spanish list but not the Asian list 1.4% of the time, and on neither list 51.6% of the time at these levels of sensitivity and specificity under the assumptions stated earlier.

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

	On Asian Surname List	On Spanish but not Asian Surname List	On neither surname list
Truly Asian	.470	.014	.516
Truly Hispanic	.005	.771	.224
Truly Black or NW White	.005	.027	.968

Illustration of the Bayesian Algorithm's "Updating"

The Bayesian Algorithm takes geocoded race/ethnic distributions a multinomial vector of Bayesian prior probabilities and updates them with the joint surname test result to produce a multinomial vector of posterior probabilities. These posterior probabilities represent the update probability of an individual with a given surname test result in a given Census block group (as represented by its race/ethnic composition) of being each of the four race/ethnicities. Three of these probabilities (all but white, for example), can be entered into a regression direct as predictors to estimate race/ethnic disparities.

Consider three hypothetical Census block groups. Block Group A is 25% each Black, Hispanic, Asian, and White. Block Group B is 50% Hispanic, 20% Black, 20% White, and 10% Asian. Block Group C is 67% White and 11% each Black, Hispanic and Asian. Tables 2a, 2b, and 2c illustrate the posterior probabilities of the Bayesian Algorithm under the current surname list sensitivities and specificities for the priors implied by Block Groups A, B, and C, respectively.

Table 2a. Posterior Probabilities for Block Group A

Race	On Asian Surname List	On Spanish Surname List ONLY	On Neither Surname List
Asian	**0.971**	0.017	0.193
Hispanic	0.010	**0.920**	0.084
Black	0.010	0.032	**0.362**
White/Other	0.010	0.032	**0.362**

Table 2b. Posterior Probabilities for Block Group B

Race	On Asian Surname List	On Spanish Surname List ONLY	On Neither Surname List
Asian	**0.919**	0.004	0.094
Hispanic	0.045	**0.970**	0.203
Black	0.018	0.013	**0.352**
White/Other	0.018	0.013	**0.352**

Table 2c. Posterior Probabilities for Block Group C

Race	On Asian Surname List	On Spanish Surname List ONLY	On Neither Surname List
Asian	**0.926**	0.014	0.068
Hispanic	0.009	**0.793**	0.029
Black	0.009	0.027	0.127
White/Other	0.056	0.166	**0.775**

For Block Group A, with even priors across the four groups, appearance on either surname list results in a 92+% chance of being in the corresponding group (higher for Asians because of the slightly greater specificity). Those on neither list are equally likely to be Black or White (36%), but with a non-trivial chance of being Asian (19%), because of the relatively low sensitivity of the Asian list.

Block Group B is majority Hispanic. Its posterior probabilities are very similar to Block Group A's. Appearing on either surname list still confers a 92+% chance of being in the corresponding group, but the probability of being Hispanic is about 5% higher and the probability of being Asian about 5% lower in each case. For those appearing on neither list, Hispanic becomes third most likely after Black and White.

Block group C is predominantly White. Here appearance on the Asian surname list, Spanish surname list, and neither list correspond strongly to posterior probabilities of being Asian, Hispanic, and White, respectively, but with Hispanic and White

probabilities in these cases only at 78-79%. Here White becomes the second most likely classification for those appearing on one of the surname lists.

Classification

As will be seen below, the Bayesian Algorithm performed better than each of the alternatives by both measures of classification performance. Which of the alternatives is next best differs between the two measures of classification performance.

Correlation with True Race/Ethnicity

Table 3a below displays the correlation with self-reported race/ethnicity for each of the three methods for each of the four race/ethnic groups. All reported correlations are statistically significant at $p<0.05$. The Bayesian Algorithm correlates at 0.62 or higher with all four indicators of race/ethnicity. Sequential Classification is the next best by this measure, with similar performance for Hispanics and Asians, somewhat lower performance for whites, and notably lower performance for blacks. Geocoding alone was near the performance of the Bayesian algorithm and notably better than Sequential Classification for blacks, but performed less well than the other two algorithms for all other groups, performing especially poorly for Hispanics. As a rough approximation, one can interpret 1-R^2 as the proportion information lost (and R^2 as the proportion of information retained) when imputing race/ethnicity with correlation R. Examining the overall weighted averages suggests that the Bayesian Algorithm contains about 45% of the information of self-reported race/ethnicity (so that a sample of 1000 without self-report) would contain about as much information as 450 cases with self-reported race/ethnicity with respect to classification. The Bayesian Algorithm retains more than twice as much information as geo-coding alone and about one-quarter more information than Sequential Classification.

Preliminary results with additional data are displayed in Table 3b. The same general patterns hold here as before, except that overall performance is improved for all methods, especially for Hispanics. Preliminary analyses by region (West vs. All Other) in the larger data set (N=9,991) found similar performance across regions except poorer performance for blacks in the West. This poorer performance where blacks were less common affected the Bayesian Algorithm the least and affected Sequential Classification the most.

Table 3a: Correlation of Algorithm Output with Self-Reported Race/Ethnicity (one plan, n=1,821)

	Correlation with Self-Reported Race/Ethnicity				Weighted Average
	Hispanic	Asian	Black	White/Other	
Bayes	0.67	0.62	0.62	0.67	0.66
Geocode	0.26	0.44	0.58	0.48	0.46
Sequential Classification	0.67	0.59	0.47	0.60	0.59

Table 3b: Preliminary Correlation of Algorithm Output with Self-Reported

Race/Ethnicity (two plans, n=9,991)

	Correlation with Self-Reported Race/Ethnicity				Weighted Average
	Hispanic	Asian	Black	White/Other	
Bayes	0.77	0.62	0.67	0.69	0.71
Geocode	0.47	0.36	0.64	0.53	0.54
Sequential Classification	0.76	0.62	0.54	0.63	0.65

Bias of Classification (Departures from True Proportions)

Table 4 below 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. 95% margins of error are about 1-2 percentage points in this table, so that differences greater than that are likely to represent bias. Differences of more than 2-3 percentage points between methods are likely to represent significantly different bias for the methods.

The overall RMSE of error is also displayed for each method. Its margin of error is about 1 percentage point, with a margin of error for differences between methods of no more than 1 ½ percentage points. The Geocoding Algorithm substantially overestimates the prevalence of Hispanics, moderately overestimates the prevalence of Asians, and slightly underestimates the prevalence of blacks and whites. This means that within these zip codes, Hispanics and Asians are less likely to be members of these health plans, perhaps as a consequence of health insurance status. Sequential Classification is very accurate for Hispanics, as it is not influenced by this selection, but it has poor sensitivity for Asians (underestimating their prevalence by near a factor of two) and very poor sensitivity for blacks (underestimating their prevalence by nearly a factor of three). This results in too plan members being classified as white (especially blacks and Asians).[3] The Bayesian method is the most accurate overall, with an average systematic error of 3.8%, followed by 6.2% for Geocoding and 9.9% for Sequential Classification. It slightly overestimates Hispanic and Asian prevalence, influenced by lower rates of plan membership in these groups.

Preliminary analyses on the set of 9,991 cases showed a similar pattern (results not shown). The RMSE for the Bayesian and Geocoding methods decreased (as expected) with increased sample size, suggesting the statistical property of consistency, whereas the RMSE of Sequential Classification did not, suggesting that bias dominates the RMSE.

Table 4: Accuracy of Overall Estimates of Race/Ethnic Prevalence

	Estimated Percentage in Each Group	Weighted Average Overall RMSE

[3] In particular, more segregated blacks will be classified as black, along with those who live near segregated blacks. Less segregated blacks will tend to be misclassified as white. To the extent that racial/ethnic patterns correlated with SES and that SES correlates with health disparities, this may result in systematic errors (bias) in estimates of disparities.

	Hispanic	Asian	Black	White/Other	
SELF-REPORT	11.1%	10.5%	10.2%	68.2%	(0)
Bayes	13.8%	12.7%	9.6%	63.9%	3.8%
Geocode	16.1%	14.0%	8.8%	61.1%	6.2%
Sequential Classification	11.4%	5.3%	3.5%	79.7%	9.9%

Estimates of Disparities

Table 5 below presents the performance of the three algorithms in estimating three racial/ethnic disparities (black vs. non-Hispanic white, Hispanic vs. non-Hispanic white, Asian vs. non-Hispanic white) for each of six dichotomous health outcomes, averaged across the six outcomes. The Bayes method performed well overall, with an average error of 3 percentage points in disparities across the six measures. This is a moderate improvement upon the Geocode Algorithm (22% less Mean Squared Error) and a substantial improvement over Sequential Classification (less than 1/3 as much Mean Squared Error). The Bayes method works especially well for estimating black vs. white and Hispanic vs. white disparities, average an error of less than two percentage points, both large improvements on the other two methods. Its performance is not as strong for Asian vs. White disparities. This appears in part due to underestimation of Asian health outcomes (data not shown). In this sample, Asians appear to be less likely to have health plan membership (and perhaps health insurance in general) than others in the same Census block groups (data not shown). Errors in surname classification of Asian are correlated with acculturation and marital status, which may in turn correlate with health performance measures.

Although Sequential Classification performs reasonably well in overall classification (correlation), it performs poorly in estimating health disparities, especially black vs. white disparities. This is probably a combination of bias in classification and loss of information through discretization. In particular, this method consistently overestimates health outcomes for blacks and Hispanics (data not shown).

Of the 18 individual disparity estimates (six outcomes for each of three comparisons), the Bayes algorithm had the lowest error for 10, the Geocoding Algorithm for 5, and Sequential Classification for 3 (data not shown).

Table 5: Mean Root MSE for Health Disparity Estimates (n=1,821)

	Mean RMSE for Disparity Estimate, averaged across 6 Dichotomous Outcomes			Average of 3 Disparity Estimates
	Black vs. White	Hispanic vs. White	Asian vs. White	
Bayes	1.4%	1.8%	5.7%	3.0%
Geocode	2.7%	3.7%	3.9%	3.4%
Sequential Classification	8.2%	3.4%	4.7%	5.4%

CONCLUSIONS

Geocoding and surname analysis are most appropriately used in combination, as their respective advantages and limitations tend to offset each other (see Fremont et al, forthcoming, Table 3). Specifically, geocoding is more reliable for inferring black race while surname analysis is better for inferring Hispanic or Asian ethnicity. Together the two methods represent a reasonable approach to inferring race/ethnicity among plan members. The Bayesian Algorithm outlined here appears to be a particularly useful means of integrating these sources of information and substantially outperforms other seemingly reasonable means of combining this information. This technique performs well when race/ethnic classification is the only goal or when estimated race/ethnicity is to be a predictor in regression or other models. Thus the applications are not limited to estimation of disparities or to health applications. There is preliminary evidence that the Bayesian Algorithm and the other methods discussed perform better when the prevalence of the smaller race/ethnic groups is greater.

Advantages of this approach include the fact that it is readily implemented, reasonably accurate, serves as a basis for action, and can incorporate other aspects of context into analyses or interventions (e.g. income, pharmacy availability). Limitations include that it is not accurate enough to support individual level interventions and requires large sample sizes for good precision, since there is some inherent loss of information compared to self-reported race/ethnicity for a sample of the same size

Combined geocoding and surname analysis provides health plans a timely means to infer race/ethnicity among their plan members for the purpose of assessing disparities in health care processes and outcomes. Although self-report represents the gold standard, indirect methods (suitably validated for a sample of plan members) offer a defensible interim alternative in lieu of direct data. A combined approach can provide a viable means for assigning race and ethnicity for purposes of examining disparities in care until self-reported data can be systematically collected on all plan members.

NEXT STEPS

1. Larger sample sizes with more plans. This will improve the precision of the evaluation, allow us to estimate plan-level variation in performance, and will make the selection into the sample closer to general selection into health insurance and not plan-specific
2. Incorporate chronic condition selection when examining subgroups. Although the approach described is suited to general studies of health disparities, disease specific approaches may allow improvements for extended studies of specific conditions. Being in some disease categories (e.g. diabetic) contains race/ethnic information. If one can assume independence of disease status and surname lists (conditional on true race/ethnicity), one can easily integrate disease status into "test" for better classification, disparity estimates
3. Make the approach specific to gender and marital status. Use gender x marital status sensitivity and specificity Surname lists perform better for males than females, especially in terms of sensitivity (51% v. 43% for Asian list; 85% v. 72%

for Spanish list). Taking this into account would improve Bayesian classification and eliminate current bias in disparity estimates that may result from race by gender interactions (this may also be a factor in the problems the Bayesian algorithm is having with Asian-White disparity estimates).

4. Further evaluate and improve disparity estimates. Compare RMSE's to what would be expected from (a) an unbiased sample (b) an unbiased sample after accounting for information loss (but not correlation of classification errors with health outcomes). Further study the association between Bayesian posterior probabilities and self-report. Run disparity models with linear and quadratic posterior probability terms (and self-reported and predicted posterior terms) to detect and quantify bias in disparity estimates.

.

APPENDIX: Implementation of Bayes Algorithm

1. Let *a, b,* and *c* be the proportion of blacks, Hispanics, and Asians in the neighborhood. W=1-a-b-c= the proportion of whites
2. Let AS=1 if the name appears on an Asian surname list and AS=0 if not.
3. Let HS=1 if the name appears on a Hispanic surname list and HS=0 if not.
4. We will give precedence to the Asian list, so that we will reset HS=0 if AS=1. This results in a trinomial joint test (three mutually exclusive outcomes: AS=1, HS=1 and AS=0, HS=AS=0)
5. Let *d* and *e* be the sensitivity and specificity, respectively of the Asian List (published, ideally normed on the region. Since these are known to vary by gender and marital status, ideally one would use values that are specific to the demographics of the individual in question). To be clear, the sensitivity of the Asian Surname List is P(AS=1|Asian) and the specificity of the Asian Surname List if P(AS=0|Not Asian). These terms are defined analogously for the Hispanic Surname List.
6. Let *f* and *g* be the sensitivity and specificity, respectively of the Hispanic List (published, ideally normed on the region)
7. We will assume specificity does not vary by (incorrect) race/ethnicity.
8. In this step we convert the sensitivities and specificities of the two surname lists tests into the sensitivities and specificities of the joint (3-level) surname test outcome, as shown in Table A.1.

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

	AS=1	HS=1 AS=0	AS=HS=0
Truly Asian	d	(1-g)(1-d)	g(1-d)
Truly Hispanic	1-e	ef	e(1-f)
Truly Black or NW White	1-e	e(1-g)	eg

I

9. Apply Bayes' Theorem to Update probabilities a,b,c to posterior probabilities a1, b1, c1. Table A.2 contains the updated posterior probabilities.

Table A.2 Posterior Probabilities of Group Membership by Test Outcomes

	AS=1	HS=1 AS=0	AS=HS=0
Asian	cd/((1-c)(1-e)+cd))	c(1g)(1-d)/(bef+c(-g)(1-d)+(1-b-c)e(1-g))	cg(1-d)/ ((1-b-c)(eg)+be(1-f)+cg(1-d))
Hispanic	b(1-e)/((1-c)(1-e)+cd))	bef/(bef+c(1-g)(1-d)+(1-b-c)e(1-g))	be(1-f)/ ((1-b-c)(eg)+be(1-f)+cg(1-d))
Black	a(1-e)/((1-c)(1-e)+cd))	(ae(1-g))/(bef+c(1-g)(1-d)+(1-b-c)e(1-g))	aeg/((1-b-c)(eg)+be(1-f)+cg(1-d))
NW White	w(1e)/((1c)(1e)+cd)	w(e(1-g)/ (bef+c(1-g)(1-d)+(1-b-c)e(1-g))	weg/((1-bc)(eg+be(1f)+cg(1-d))

Note A1+B1+C1+W1 MUST=0

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: pp. 383-398.

Blustone. 1994.

Clark, William A. V. and Peter A. Morrison. 2005. "Evaluating Evidence of Discrimination in Multi-Ethnic Housing Markets," presented at 2005 annual Population Association of America meetings.

Falkenstein, Matthew R. 2002. "The Asian and Pacific Islander Surname List: As Developed from Census 2000," paper presented at the Joint Statistical Meetings, August 11, 2002.

Fremont and Lurie. 2004

Fremont, Allen M. et al. 2005. Racial and Socioeconomic Disparities in the Quality of Cardiovascular and Diabetes Care in Managed Care. *Health Affairs.*

Fremont et al. [??] . Forthcoming. USE OF GEOCODING AND SURNAME ANALYSIS TO ESTIMATE RACE AND ETHNICITY

Institute of Medicine. 2002. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*

Kestenbaum, Bert, B. Renee Ferguson, Irma Elo, and Cassio Turra. 2000. "Hispanic Identification," paper presented at the 2000 Southern Demographic Association meetings.

Kressin, et al. 2003.

Lauderdale, Diane and Bert Kestenbaum. 2000. "Asian American Ethnic Identification by Surname," Population and Development Review 19 (3), pp. 283-300.

Morrison, Peter A., et al. 2001. "Using First Names to Estimate Racial Proportions in Populations," presented at the 2001 Population Association of America meetings.

Morrison, Peter A. 2003a. "Confronting a Race-Based School Admissions Policy," Chance 16(1), 2003.

Morrison, Peter A., et al. 2003b. "Developing an Arab-American Surname List: Potential Demographic and Health Research Applications," at 2003 Southern Demographic Association meetings.

Morrison, Peter A. Forthcoming. "Lingering Effects of Discrimination: Tracing Persistence Over Time in Local Populations," *Population Research & Policy Review.*

National Quality Forum. 2002.

Perez-Stable, et al. 1995.

Passel, J. S. and D. L. Word. 1980. "Constructing the List of Spanish Surnames for the 1980 Census: An Application of Bayes' Theorem," presented at the annual Population Association of America meetings, Denver, April 10-12.

Perkins, R. C. 1993. "Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results. U.S. Census Bureau, Population Division, Technical Working Paper No. 4.

Rosenwaike & Bradshaw. 1988.

Word & Perkins. 1996.

Working Group on Quality. 2004.