Published in final edited form as:

Med Care. 2010 June ; 48(6): 540–546. doi:10.1097/MLR.0b013e3181d559e9.

Using Name Lists to Infer Asian Racial/Ethnic Subgroups in the Healthcare Setting

Eric C. Wong, MS*, Latha P. Palaniappan, MD, MS*, and Diane S. Lauderdale, PhD†

*Palo Alto Medical Foundation Research Institute, Palo Alto, CA

†Department of Health Studies, University of Chicago, Chicago, IL.

Abstract

Background—Many clinical data sources used to assess health disparities lack Asian subgroup information, but do include patient names.

Objective—This project validates Asian surname and given name lists for identifying Asian subgroups (Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese) in clinical records.

Subjects—We used 205,000 electronic medical records from the Palo Alto Medical Foundation, a multipayer, outpatient healthcare organization in Northern California, containing patient self-identified race/ethnicity information.

Research Design—Name lists were used to infer racial/ethnic subgroup for patients with self-identified race/ethnicity data. Using self-identification as the “gold standard,” sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of classification by name were calculated. Clinical outcomes (obesity and hypertension) were compared for name-identified versus self-identified racial/ethnic groups.

Results—With classification using surname and given name, the overall sensitivities ranged from 0.45 to 0.76 for the 6 racial/ethnic groups when no race data are available, and 0.40 to 0.79 when the broad racial classification of “Asian” is known. Specificities ranged from 0.99 to 1.00. PPV and NPV depended on the prevalence of Asians in the population. The lists performed better for men than women and better for persons aged 65 and older. Clinical outcomes were very similar for name-identified and self-identified racial/ethnic groups.

Conclusions—In a clinical setting with a high prevalence of Asian Americans, name-identified and self-identified racial/ethnic groups had similar clinical characteristics. Asian name lists may be a valid substitute for identifying Asian subgroups when self-identification is unavailable.

Keywords

Asian race/ethnicity; surname analysis; racial disparities

Disparities research often uses clinical data, such as Medicare and Medicaid claims records, electronic health records, hospital claims, and health plan administrative files to compare disease and healthcare use across different racial and ethnic groups. However, comparisons for specific Asian American racial/ethnic groups (eg, Korean, Vietnamese) are rarely reported. This is not because researchers are unaware of differences between Asian racial/ethnic groups, but rather because only a few health data sources (eg, birth and death

certificates) include subgroup-level information for Asians. Most clinical data sources that have race data use 4 to 6 racial/ethnic categories and include an aggregated Asian (or Asian and Pacific Islander) category. Some clinical data sources have no racial/ethnic information at all. However, interest has grown in recent years in disaggregating Asian Americans into specific racial/ethnic subgroups both because of rapidly growing subgroup populations and because of rising awareness of differences across the subgroups regarding economic status, immigration history, and risk of specific infectious and noninfectious diseases.¹⁻⁵

With limited availability of Asian racial/ethnic subgroup information, one alternative is to infer race/ethnicity on the basis of an individual's name. The Census Bureau's 1980 and 1990 lists of Spanish names are a model for the utility of name classification since they have been used for decades to infer Hispanic ethnicity in data situations where names are available but racial/ethnic information is not, in both primary data collection and in secondary analyses of registry and administrative records.⁶⁻⁸ The basic idea behind name classification is that there are names that are unlikely to occur among persons not in a specific racial/ethnic group and those names can be used to classify persons with high probability of belonging to the target racial/ethnic groups. While it is unlikely that all of the persons in a racial/ethnic group are identifiable by name, an assumption is made that those with more distinctive names have similar characteristics to all persons from the racial/ethnic group. Although this assumption has been tested by comparing demographic characteristics, to our knowledge, this has never been tested with health data.^{9,10}

Lauderdale and Kestenbaum developed surname lists for the 6 largest Asian American populations (Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese) and published an article describing their derivation in 2000.¹¹ Given name lists were created but not similarly validated.¹² To date, there has been no validation of the given name lists or of the combined use of the surname and given name lists for racial/ethnic classification. Nor has there been a test of whether a racial/ethnic group defined by name has similar health characteristics to a racial/ethnic group defined by self-identification, which is the key question when considering using name identification for disparities research.

In this project, we tested the surname and given name lists in the administrative records of a healthcare organization with a large Asian representation. The project capitalizes on recently added self-identified racial/ethnic data in patient medical records, providing a "gold standard" for validating name list classification. Specifically, the analysis addresses these questions: How complete and accurate are name lists in classifying Asians into subgroups? Is there appreciable improvement in classification when given names are used along-side surnames? How predictive is name classification in populations with varying prevalences of Asian Americans? Are health characteristics (specifically, obesity, and hypertension) similar for name-identified and self-identified racial/ethnic cohorts?

METHODS

Study Population

The Palo Alto Medical Foundation (PAMF) is a mixedpayer, outpatient, healthcare organization in the San Francisco bay area of California. The PAMF has 28 clinics and in 2007 performed nearly 2 million patient visits for approximately 480,000 active patients. In May 2008, the PAMF began collecting self-identified race/ethnicity through a one time questionnaire presented to patients during clinic office visits. The questionnaire and data collection process has been described elsewhere.¹³ Briefly, the questionnaire asks 5 questions modeled after the Census 2000 on race, Hispanic origin, ancestry, spoken language, and need for interpreter.¹⁴ The responses to the questions are both categorical (race, Hispanic origin, need for interpreter) and free-text (ancestry, language). Patients may

identify with up to 2 races, and up to 2 ancestries, and all other questions are single responses. By July 2009, this information had been collected on approximately 205,000 patients.

The study population included all patients who had a clinic visit between May 2008 and July 2009, when a data file was created for these analyses. Every patient record was assigned several inferred race/ethnicity codes (using different classification algorithms involving patient given names and surnames) and these were compared with self-identified race/ethnicity. No patients were contacted for this study, only de-identified data in existing electronic health records were used. The protocol was approved by the PAMF Institutional Review Board.

Self-Identified Racial/Ethnic Identification

Responses to the race, Hispanic origin, and ancestry questions on the new clinic questionnaire were used to create self-identified race/ethnicity. Some aggregation was necessary to digest these responses for race, Hispanic origin, and ancestry into a single race/ethnicity code. Only single-race and agreeing responses were classified as an Asian subgroup. For example, a response of race = “Chinese,” Hispanic origin = “Not Hispanic/Latino,” and ancestry = “Chinese” would be considered single-race, agreeing and classified as “Chinese.” Multiracial, nonagreeing, and other responses were all classified accordingly, and were not counted towards the “gold-standard” racial/ethnic classification of Asian subgroups.

Inferred Racial/Ethnic Classification by Name Lists

The Lauderdale and Kestenbaum name lists were matched to the surname (also known as last or family name) and given name (also known as first or Christian name) on each patient record (the name lists are available to other researchers from Diane Lauderdale at lauderdale@uchicago.edu). The surname lists were developed for 2 data situations: one where no race information is available (unconditional) and the other when the broad category of “Asian” (or “Asian/Pacific Islander”) is known but subgroup information is absent (conditional). Some names are only predictive of a specific Asian racial/ethnic subgroup if the person is known to be Asian (conditional). For example, the surname “Bang” is quite specific to Koreans if the person is known to be Asian, but without any racial/ethnic information, is more likely to be non-Asian. The given name lists are all unconditional, meaning they do not depend on known race information. Name lists are identified for 6 racial/ethnic subgroups: Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese.

Each surname and given name was noted as included in a name list or not. The surname could be absent from all of the surname lists, only on a conditional list, or on both a conditional and unconditional list. Surnames on a conditional list are only considered identifying when the self-reported responses included any of the Asian categories (emulating the response from a questionnaire that only offered a single pan-Asian race category). The given name could be on a given name list or not. Each patient record was assigned 4 inferred race/ethnicity codes based on 4 different name matching algorithms: unconditional surname only; conditional surname only (value assigned only if the self-identified race/ethnicity was any Asian subgroup); EITHER unconditional surname OR given name (assignment when either is Asian), and EITHER conditional surname OR given name. For example, a patient with the name “Hee-Kyung Bang” would be Korean for 3 of the 4 algorithms since “Hee-Kyung” is on the Korean given name list and “Bang” is on the conditional Korean surname list, but not the unconditional Korean surname list. Multiracial, nonagreeing surname and given name classification, or other responses were all classified accordingly, and were not

counted towards “true” racial/ethnic classification of Asian subgroups. We also considered algorithms that required matching of both surname and given name, ie, classification only when both surname and given name were members of a list, but they identified so few patients that we omitted those algorithms from this presentation.

We also matched patient surnames to the 1990 Census Spanish surname list,¹⁵ allowing us to compare Asian name list performance to Spanish surname list performance. Patients with Asian given names and Spanish surnames were classified as discordant unless the given name was Filipino, in which case they were classified as Filipino (since many Filipino surnames are Spanish). In secondary analyses (not shown), individuals with Filipino given names and Spanish surnames were classified like other Asian subgroups as discordant, but the accuracy of the Filipino classification decreased.

Analysis

Our first analyses compared 4 name-based inferred race/ethnicity algorithms to self-identified race/ethnicity. The 4 algorithms were (1) unconditional surname; (2) conditional surname; (3) EITHER unconditional surname OR given name; and (4) EITHER conditional surname OR given name. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. We examined whether name list performance differed by sex or age (aged 65 and older vs. younger). The sensitivity and specificity were examined with 95% binomial confidence intervals, and PPV and NPV were examined with 95% confidence intervals using Mercaldo’s standard logit and adjusted method.¹⁶ Since tight precision was observed (interval lengths ranged: 0.0001–0.0597) due to small variance and the large sample size, confidence intervals are not included in the tables.

Since the PPV and NPV depend on the prevalence of specific Asian subgroups in the target population, we estimated the PPV and NPV in geographic areas with varying Asian racial/ethnic concentrations: the whole United States (low concentration), the state of California (medium concentration), and the San Francisco Bay Area (the Census’s San Jose-San Francisco-Oakland CSA; high concentration). For these analyses, the algorithms using unconditional surname only and EITHER unconditional surname OR given name were presented. In addition, we calculated PPV and NPV using the conditional surname lists for all national Asian American populations. When the data source has broad Asian race information, then one can identify a population that is 100% Asian, and only the distribution of subgroups among Asians affects the PPV and NPV. For all geographic areas, the prevalence of Asians was estimated using the 2007 American Community Survey.¹⁷

Our final analyses assessed whether clinical characteristics varied similarly across name-identified racial/ethnic groups and across self-identified racial/ethnic groups. Since nearly every electronic health record at PAMF includes height, weight, and blood pressure, we compared the distribution of body mass index (BMI) categories and proportion with hypertension for adults (aged 18 and older). The BMI categories were underweight (<18.5 kg/m²), normal (18.5 – <25 kg/m²), overweight (25 – <30 kg/m²), and obese (≥ 30 kg/m²). Hypertension was defined as a systolic blood pressure greater than 140 mm Hg or a diastolic blood pressure greater than 90 mm Hg. We presented and tested whether the distribution of BMI categories differed between self-identification or name-classification, using χ^2 tests of independence at $P < 0.01$. Prevalence of hypertension was compared with tests of 2 binomial proportions. All statistical analyses were performed using SAS 9.2 (Cary, NC).

RESULTS

The PAMF study population consisted of 204,787 active patients. This cohort was 44% male and 56% female. The mean and median age was 39 years. The number of patients within each racial/ethnic group as identified by self-identification and the 4 name matching algorithms is presented in Table 1. About 91% of the population self-identified as belonging to 1 of the 6 Asian racial/ethnic subgroups, Hispanic, Black, or Non-Hispanic White. The remainder of the population was multiracial or a different race/ethnicity (8%). The name lists classified fewer patients than self identification in nearly all racial/ethnic subgroups, confirming the less-than-perfect sensitivity of the name lists. Of the race/ethnicities inferred by name algorithms, only 0.03% had nonagreeing classification by surname and given name, of which 46% were men and 54% women.

The overall sensitivities and specificities of the name lists are presented in Table 2. Compared with the Asian surname lists, the Spanish surname list had intermediate sensitivity, but lower specificity. The “EITHER” algorithms, which also considered given names, had higher sensitivities and specificities compared with surname only. Across the Asian subgroups, specificities were uniformly high; there was more variability among sensitivities.

When we stratified the analyses by sex, sensitivities were generally higher for men than women, while specificities were uniformly high for both sexes (see data in Appendix, online only, available at:

<http://www.editorialmanager.com/mdc/download.aspx?id=82993&guid=7f29ef6b-139c-48c8-a71e-0ad99e74f3ef&scheme=1>). The difference by sex in sensitivity was largest for Japanese. When we stratified by age 65 years and older versus younger, the sensitivities were greater for older persons, particularly when identification was by either surname or given name. Specificities were uniformly high (Appendix, online only, available at: <http://www.editorialmanager.com/mdc/download.aspx?id=82993&guid=7f29ef6b-139c-48c8-a71e-0ad99e74f3ef&scheme=1>).

Using the sensitivity and specificity determined from our PAMF sample, PPV and NPV were estimated for geographic areas with different concentrations of Asian subgroups (Table 3). The national distribution had PPVs ranging from 0.31 to 0.61 and NPVs ranging from 0.99 to 1.00. For areas with high concentration of Asians, such as the San Francisco Bay Area, the PPVs increased to 0.63–0.89 and NPVs only slightly decreased to 0.96–1.00. Within the PAMF sample, the PPVs and NPVs ranged from 0.71 to 0.93 and 0.89 to 1.00, respectively. And for the example of a known, all-Asian group, the PPVs and NPVs ranged from 0.94 to 0.98 and 0.88 to 0.99, respectively. Although PPVs and NPVs necessarily vary with prevalence, geographic areas with distributions similar to the San Francisco Bay Area would have high PPV. The PPV and NPV are very high when Asian race data are available and the conditional surname list can be used.

Our final analyses assessed whether clinical characteristics varied similarly across name-identified racial/ethnic groups and across self-identified racial/ethnic groups. For these analyses we used the algorithms that classify by either surname or given name, since these algorithms consistently outperformed the surname only lists. Figure 1 shows the distribution of BMI categories for each Asian racial/ethnic subgroup, comparing the distributions classified by name lists versus self-identified race/ethnicity. No subgroup had a statistically significant difference in BMI distribution between self-identification and name-based (unconditional or conditional) classification. Similarly, there was no significant difference in

the prevalence of hypertension between groups that were self-identified and name-identified (results not shown).

DISCUSSION

We evaluated Asian racial/ethnic group identification inferred from surname and given name lists for 6 Asian American racial/ethnic groups (Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese) against self-identified race/ethnicity, using electronic medical records from a large healthcare organization. We have 2 key findings. First, name identification for Asian subgroups is more complete and no less accurate when given names are considered in addition to surnames. Second, health characteristics are very similar between self-identified and name-classified Asian racial/ethnic subgroups.

Comparing different name-based algorithms using surnames alone and a combination of surnames and given names, we found that the completeness of the name lists as measured by sensitivity was moderate and the accuracy as measured by specificity was very high. The racial/ethnic subgroups had a spectrum of sensitivities (range: 0.45–0.78), with higher sensitivity for Vietnamese, Japanese, and Chinese. Filipinos had the lowest sensitivity. Specificities ranged from 0.97 to 1.00, meaning that names not associated with Asian individuals were not on the name lists, which was expected, given how the name lists were initially built. The Spanish surname list, a validated name list that has been frequently used in health research, had intermediate sensitivity compared with the Asian name lists, but lower specificity.

We found that much was gained by combining given and surname lists. For example, the sensitivities of some subgroups nearly doubled when adding given names (eg, Asian Indians). As we expected, the name lists had higher sensitivities and specificities for men than women likely due to name changes following marriage. We found that sensitivities were also higher for persons aged 65 and older, suggesting more within-group marriage for older women and more distinctive given names among men and women because of foreign birth.

Since the predictive values depend on the concentration of Asians in the population, we compared PPV and NPV for locations with different prevalences of Asian racial/ethnic subgroups. For locations with low Asian prevalence, the lists had moderate PPV (range: 0.31–0.56). However, as the concentration of Asians increased, the PPV rapidly rose. Locations with racial/ethnic subgroup prevalence similar to those of the state of California or the San Francisco Bay Area may be reasonably confident that persons identified as Asian by the name lists are highly likely to be Asian (SF Bay Area PPV: 0.70–0.87). When name identification can be conditioned on the broad race category for Asians, the PPV is very high for all Asian subgroups.

Most importantly, we found that health characteristics for specific racial/ethnic groups were very similar when identification was by name list and by self-identification. This was true for both health outcomes we tested: obesity and hypertension.

Few attempts have been made to evaluate the performance of Asian name lists. Lauderdale and Kestenbaum compared the performance of their surname lists, built from US Social Security files, on a separate evaluation data set from the Census.¹¹ The sensitivity of the Asian (unconditional) surname lists ranged from 0.24 to 0.68 with a PPV of 0.83 to 0.93. Quan et al validated Lauderdale's Chinese surname list against a Canadian national health survey and found a sensitivity of 0.53 and PPV of 0.92 in a Canadian population with 1.6% Chinese.¹⁸ Our validation of the same Chinese surname list in a US population had higher sensitivity, but lower PPV. There are a few commercial hybrid methods that combine

geocoding and surname analysis that have been validated on UK populations, but not on a US population,¹⁹ and new Bayesian surname and geocode methods demonstrate potential.²⁰ To our knowledge, there has been no prior evaluation of combining given name and surname classification, and no prior study that compared health characteristics of name-identified and self-identified groups.

Our study has both strengths and limitations. The PAMF's service area spans over many of the San Francisco Bay Area counties. The San Francisco Bay Area, one the most ethnically diverse metropolitan areas, is home to the largest population of Asians in the nation. The large underlying Asian population guarantees adequate representation of each Asian racial/ethnic subgroup, and makes it an ideal place to determine sensitivity and specificity of name lists in a real world application. However, few areas of the country have as high Asian prevalence as the San Francisco Bay Area and name identification will be less accurate in communities with more typical racial/ethnic distributions. Additionally, this study capitalized on PAMF's electronic medical record system. While all clinical data sources include patient names, access to identifiable records is often difficult or complicated for researchers. Another limitation of name list identification for Asian subgroups is that sensitivity varies across racial/ethnic groups, for reasons related to name characteristics. Asian Indian name identification is less complete because there is a large universe of names used in the linguistically heterogeneous South Asian subcontinent. The original name lists omitted names that had fewer than 5 occurrences in the derivation files at the Social Security Administration (for confidentiality reasons), and that reduced the sensitivity of the Asian Indian list. Finally, Korean surnames have a unique problem. A small number of surnames are extremely common in Korea, but some of them also occur among non-Asian populations (eg, Lee) and others are Chinese in origin (eg, Chang). These names are not specific enough to Koreans and cannot be used for name identification. However, our comparisons of clinical outcomes demonstrate that even though some Asian subgroups are less completely identifiable by name, they appear to be just as representative of the entire group as those groups with more complete name-identification, such as Vietnamese.

This paper has shown that when clinical data sources have names but limited or no Asian race/ethnicity data, name lists may be used to infer specific Asian racial/ethnic subgroups. In these situations, clinicians and decision makers could use name lists to identify potential racial/ethnic disparities in disease or in healthcare receipt, or target specific populations to provide more culturally competent care. Using the inverse of the sensitivity estimates from this study as sample weights, the number of people identified by surname and given name can be adjusted to estimate the actual racial/ethnic population size.

We make several recommendations to users of these Asian subgroup name lists. First, organizations planning to use name lists to infer Asian subgroups should consider using the given name list together with the surname lists. For situations when the broad category of "Asian" is known, the combination of conditional surname and given name should be used with the known race information. Second, one should be aware of the differences in sensitivity across subgroups and sex when applying name lists to a target population. Name list identification is also more accurate and complete for older populations. Third, one should be attentive to the prevalence of Asian subgroups in the target population. Even though the specificities of the lists are high, the accuracy of name lists as measured through the PPV and NPV vary dramatically by the concentration of Asians in the target population. Finally, we hope our findings will lead to new studies of racial/ethnic health and healthcare disparities in areas with substantial concentrations of Asian such as California, the San Francisco Bay Area, Los Angeles, New York, and Hawaii and in data sources where there is Asian race information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Jessica Shin and Ariel Holland for their assistance in manuscript preparation and submission.

Supported by funds from National Institutes of Diabetes, Digestive and Kidney Diseases (1 R01 DK081371–01A1 Identifying Disparities in Type 2 Diabetes Among Asian Americans: The Pan Asian Cohort Study) from the period January 15, 2009–December 31, 2013 (to L.P.P.); and also from American Heart Association (0885049N–Asian American Heart Study) from the period July 1, 2008–June 30, 2010 (to L.P.P.).

REFERENCES

1. Hsiao AF, Wong MD, Goldstein MS, et al. Complementary and alternative medicine use among Asian-American subgroups: prevalence, predictors, and lack of relationship to acculturation and access to conventional health care. *J Altern Complement Med*. 2006; 12:1003–1010. [PubMed: 17212572]
2. Lauderdale DS, Rathouz PJ. Body mass index in a US national sample of Asian Americans: effects of nativity, years since immigration and socioeconomic status. *Int J Obes Relat Metab Disord*. 2000; 24:1188–1194. [PubMed: 11033989]
3. Chen MS Jr. Cancer health disparities among Asian Americans: what we do and what we need to do. *Cancer*. 2005; 104:2895–2902. [PubMed: 16270313]
4. Srinivasan S, Guillermo T. Toward improved health: disaggregating Asian American and Native Hawaiian/Pacific Islander data. *Am J Public Health*. 2000; 90:1731–1734. [PubMed: 11076241]
5. Wong E, Lauderdale D, Fortmann S, et al. Heterogeneity in cardiovascular risk factors among Asian-American subgroups. *Circulation*. 2008; 117:e198–e291.
6. Miller JE, Guarnaccia PJ, Fasina A. AIDS knowledge among Latinos: the roles of language, culture, and socioeconomic status. *J Immigr Health*. 2002; 4:63–72. [PubMed: 16228761]
7. Morgan RO, Wei II, Virnig BA. Improving identification of Hispanic males in Medicare: use of surname matching. *Med Care*. 2004; 42:810–816. [PubMed: 15258483]
8. Polednak AP. Estimating cervical cancer incidence in the Hispanic population of Connecticut by use of surnames. *Cancer*. 1993; 71:3560–3564. [PubMed: 8490904]
9. Rosenwaike I. Surname analysis as a means of estimating minority elderly: an application using Asian surnames. *Res Aging*. 1994; 16:212–227.
10. Shin EH, Yu EY. Use of surnames in ethnic research: the case of Kims in the Korean-American population. *Demography*. 1984; 21:347–360. [PubMed: 6479394]
11. Lauderdale D, Kestenbaum B. Asian American ethnic identification by surname. *Popul Res Policy Rev*. 2000; 19:283–300.
12. Lauderdale DS, Kestenbaum B. Mortality rates of elderly Asian American populations based on Medicare and Social Security data. *Demography*. 2002; 39:529–540. [PubMed: 12205756]
13. Palaniappan L, Wong E, Shin J, et al. Collecting patient race/ethnicity and primary language data in ambulatory care settings: a case study in methodology. *Health Serv Res*. 2009; 44:1750–1761. [PubMed: 19555396]
14. United States Census 2000. Long Form Questionnaire—Form D-2. Vol. 3. US Department of Commerce; Bureau of the Census; 2000. p. 12
15. Word, D.; Perkins, R. Building a Spanish Surname List for the 1990's—A new approach to an old problem. US Census Bureau; Washington, DC: 1996.
16. Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case-control studies. *Stat Med*. 2007; 26:2170–2183. [PubMed: 16927452]
17. American Community Survey. Asian Alone by Selected Groups Universe: total Asian Alone population. US Census Bureau; 2007. p. C02006
18. Quan H, Wang F, Schopflocher D, et al. Development and validation of a surname list to define Chinese ethnicity. *Med Care*. 2006; 44:328–333. [PubMed: 16565633]

19. Fiscella K, Fremont AM. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Serv Res.* 2006; 41:1482–1500. [PubMed: 16899020]
20. Elliott MN, Fremont A, Morrison PA, et al. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res.* 2008; 43:1722–1736.

FIGURE 1.

Distribution of adult Body Mass Index (BMI) categories by Asian subgroup and classification method (name-inferred or self-identified). Top panel, Unconditional Either algorithms applied to the entire patient population. Bottom panel, Conditional Either algorithms applied to the patient population of all Asian Americans. For every Asian subgroup, the distribution using inferred classification is shown above, and the distribution using self-identification (self-ID) below. These 2 methods of identification/classification were not statistically significantly different in any subgroup by χ^2 tests.

TABLE 1
Counts of Patients Within Each Racial/Ethnic Subgroup as Identified by Self-Identification or Name Lists

	Self-Identified	Unconditional Surname		Conditional Surname	
		Surname Only	Either Surname or Given Name	Surname Only	Either Surname or Given Name
Asian Indian	27,424	6,935	13,301	11,796	16,225
Chinese	23,598	16,439	17,798	17,674	18,365
Filipino	7,060	2,073	2,195	2,940	2,958
Japanese	2,884	2,497	3,029	2,075	2,384
Korean	2,341	1,295	1,544	1,490	1,597
Vietnamese	2,561	2,259	2,569	2,159	2,317
Hispanic	17,597	16,910	16,794	—	—
Black	3,912	—	—	—	—
Multiracial and other	19,157	—	—	—	—
Non-Hispanic white	98,253	—	—	—	—

Four name-matching algorithms were used to identify potential racial/ethnic groups: (1) unconditional surname only, (2) conditional surname only, (3) either unconditional surname or given name, and (4) either conditional surname or given name. The unconditional name lists were used when no race/ethnicity information was known. The conditional lists were used when a broad race classification of “Asian” was known.

TABLE 2

Overall Sensitivity and Specificity of Unconditional and Conditional Names Lists

	<u>Unconditional Surname</u>		<u>Conditional Surname</u>	
	Sensitivity	Specificity	Sensitivity	Specificity
Surname only				
Asian Indian	0.24	1.00	0.42	1.00
Chinese	0.63	0.99	0.73	0.99
Filipino	0.23	1.00	0.40	1.00
Japanese	0.62	1.00	0.66	1.00
Korean	0.42	1.00	0.53	1.00
Vietnamese	0.70	1.00	0.76	1.00
Hispanic	0.61	0.97	—	—
Either surname or given name				
Asian Indian	0.45	0.99	0.58	0.99
Chinese	0.67	0.99	0.75	0.99
Filipino	0.25	1.00	0.40	1.00
Japanese	0.75	1.00	0.77	1.00
Korean	0.49	1.00	0.58	1.00
Vietnamese	0.76	1.00	0.79	1.00

The unconditional name lists are used when no race/ethnicity information is known. The conditional list is used when a broad race classification of “Asian” is known. The Either combination of name lists, which use both the Asian given name and surname, perform better than just using surname alone.

TABLE 3
Examples of Positive Predictive Value (PPV) and Negative Predictive Value (NPV) of Name Lists by Geography

	USA			California			SF Bay Area			PAMF			USA Asians (Conditional)		
	Prev	PPV	NPV	Prev	PPV	NPV	Prev	PPV	NPV	Prev	PPV	NPV	Prev	PPV	NPV
Surname only															
Asian Indian	0.9%	0.44	0.99	1.3%	0.55	0.99	2.9%	0.73	0.98	13.4%	0.93	0.89	19.4%	0.95	0.88
Chinese	1.0%	0.43	1.00	3.2%	0.71	0.99	7.4%	0.85	0.97	11.5%	0.91	0.95	23.0%	0.94	0.92
Filipino	0.8%	0.46	0.99	3.2%	0.78	0.98	5.2%	0.85	0.96	3.4%	0.79	0.97	18.2%	0.98	0.88
Japanese	0.3%	0.32	1.00	0.8%	0.59	1.00	1.0%	0.64	1.00	1.4%	0.72	0.99	6.1%	0.95	0.98
Korean	0.4%	0.55	1.00	1.2%	0.76	0.99	1.0%	0.74	0.99	1.1%	0.76	0.99	10.2%	0.94	0.95
Vietnamese	0.5%	0.61	1.00	1.5%	0.83	1.00	2.6%	0.89	0.99	1.3%	0.80	1.00	11.4%	0.97	0.97
Either surname or given name															
Asian Indian		0.40	1.00		0.51	0.99		0.70	0.98		0.92	0.92		0.95	0.91
Chinese		0.40	1.00		0.68	0.99		0.84	0.97		0.89	0.96		0.94	0.93
Filipino		0.46	0.99		0.78	0.98		0.85	0.96		0.79	0.97		0.98	0.88
Japanese		0.31	1.00		0.58	1.00		0.63	1.00		0.71	1.00		0.95	0.99
Korean		0.53	1.00		0.74	0.99		0.72	0.99		0.74	0.99		0.94	0.95
Vietnamese		0.56	1.00		0.79	1.00		0.87	0.99		0.76	1.00		0.96	0.97

The positive predictive value (PPV) and negative predictive value (NPV) for populations with different prevalence (Prev) of Asian Americans: the national population (USA), the San Francisco Bay metropolitan area (San Jose-San Francisco-Oakland CSA), California, the PAMF patient population, and the subset of the national population that are Asians (USA Asians). Populations were determined from the 2007 American Community Survey. The sensitivity and specificities of the name lists as estimated in the PAMF population were used in calculating PPV and NPV. The unconditional name algorithms were used in the first 4 examples, and the conditional algorithms were used in the last example.