Population Research and Policy Review **19:** 283–300, 2000.

Asian American ethnic identification by surname

DIANE S. LAUDERDALE[1] & BERT KESTENBAUM[2]
[1]*Department of Health Studies, University of Chicago, Illinois, USA;* [2]*Office of The Chief Actuary, Social Security Administration*

Abstract. Few data sources include ethnicity-level classification for Asian Americans. However, it is often more informative to study the ethnic groups separately than to use an aggregate Asian American category, because of differences in immigration history, socioeconomic status, health, and culture. Many types of records that include surnames of persons offer the potential for inferential ethnic classification. This paper describes the development of surname lists for six major Asian American ethnic groups: Chinese, Japanese, Filipino, Korean, Asian Indian, and Vietnamese. The lists were based on Social Security Administration records that include country of birth. After they were compiled, the lists were evaluated using an independent file of census records. The surname lists have a variety of applications for researchers: identification of individuals to target for study participation; inference of ethnicity in data sources lacking ethnic detail; and characterization of the ethnic composition of a population.

Keywords: Asian Americans, Names, Ethnic groups/classification

Introduction

The Asian American population has grown rapidly over the past three decades. The result of this growth is a numerically large minority group – over 10 million persons – most of whom are foreign-born. The extension of the racial data collection system in the USA to include this population has been inconsistent. Only recently has a race category for Asian Americans been routinely included on forms. For example, before 1980, application forms for Social Security numbers simply had a category 'other' for all non-black, non-white applicants. Although race questions on forms now generally include the choices 'Asian' or 'Asian and Pacific Islander', ethnic-specific categories such as Asian Indian, Korean, or Chinese would be more useful for research purposes.

The advantage of ethnicity-level identification is that it does not mask important differences among the groups. Whereas most Japanese American adults are native-born, most adults of the other ethnic groups are foreign-born. Socioeconomic status (SES) varies markedly among ethnic groups (Barringer et al. 1993: 231–267): Japanese and Asian Indian Americans are among the

wealthiest groups in the country; Southeast Asians have on average much lower levels of education and higher levels of poverty. Because some Asian groups are socioeconomically disadvantaged compared to whites, while others are advantaged, the numerous health indicators related to SES, such as mortality, are relatively uninformative when applied to the 'average' Asian American. In fact there is remarkably little information about the basic health status of Asian American ethnic groups. *Healthy People 2000*, the report on national health objectives, states "An adequate depiction of the health of Asian and Pacific Islander Americans is constrained because data cannot be stratified by subgroups" (US Department of Health and Human Services 1991: 36).

Few data sources allow one to identify specific Asian American ethnic groups. One that does is the decennial census, which has always listed each numerically substantial Asian ethnic group as a race option, beginning with Chinese in 1860. In 1990, Asian ethnic options were for the first time grouped together under a single rubric, 'Asian or Pacific Islander'. Increasing the opportunities for ethnicity-specific analyses, the National Center for Health Statistics expanded its race code structure to include six Asian ethnic groups (Chinese, Japanese, Filipino, Korean, Vietnamese and Asian Indian) for both vital status records and the National Health Interview Survey in 1992 (Kuo & Porter 1998; Yu & Liu 1992).

However, sources used in public health and demographic research often do not include race or ethnic information or only use a general 'Asian' term. Records with names of persons offer the possibility of inferential ethnic classification. One could potentially use such inferred ethnic classification to select records by surname from an administrative database, such as the enrollment file of a health maintenance organization, and then determine rates of hospitalization, procedures, or diagnoses. One could use surnames to identify local concentrations of ethnic groups in the years between decennial censuses, or the ethnic composition of registered voters, students, or homeowners (Abrahamse et al. 1994). Surnames could serve as a means of estimating the completeness of ethnic or racial identification where the information is incompletely recorded or recorded by a third party, such as on a death certificate. One could select persons by surname from a roster or directory as a means of oversampling minority groups to participate in a cohort or panel study.

The inference of ethnicity from surname is most familiar in the United States for Spanish surnames. The Census Bureau has been developing and using Spanish surname lists since 1950 (Perkins 1993). Although the Census Bureau's lists are not the only publicly available Spanish surname tool (Buechley 1976), its two most recent products, developed in conjunction

with the 1980 and 1990 censuses (Word & Perkins 1996; Passel & Word 1980), have been widely used by researchers. There are no Asian surname lists with a similar level of acceptance or recognition. A consideration of the development of the 1990 Spanish surname list makes clear the difficulty in constructing lists of Asian surnames. The most recent Spanish surname list was compiled from a sample of 1990 census records for approximately 1.9 million heads of household and unrelated individuals (excluding ever-married females), a file created in conjunction with the 1990 post-enumeration survey. Each record contained the surname as well as responses to census questions on race and Hispanic ethnicity. About 200,000 in the sample identified themselves as Hispanic.

Even a national sample this large is inadequate for deriving surname lists for Asian ethnic groups. The total number of Asians on this census file of 1.9 million is only about 40,000. Of these less than 10,000 are of any one Asian ethnic origin. This number represents one-twentieth the size of the Hispanic sample. A file many times larger is needed to yield the needed numbers of records for persons of a specific Asian ethnic group. In the uniquely large-scale effort described here we instead turned for surname list derivation to Social Security Administration (SSA) files containing many millions of records. We derived lists for each of the six largest Asian American ethnic groups: Chinese, Filipino, Indian, Japanese, Korean and Vietnamese. We hypothesized that in data situations where there is an Asian race classification available, the race information could be used to increase both accuracy and completeness of surname-inferred ethnic identification. Therefore, we derived surname lists for two data contexts. We derived lists which make inference of ethnicity conditional on Asian race identification (conditional lists) for use when race data are available, and we derived unconditional lists for use with records which do not include race classification. We described the accuracy and completeness of the surname lists in identifying members of Asian ethnic groups in the SSA records, and we turned to the 1990 census surname file to evaluate the lists with a file quite different than the source file. For comparison, we also evaluated with the census file Asian surname lists previously developed by others.

Materials and methods

Derivation of surname lists

Deriving surname lists empirically involves using a large file of records for a population with an ethnic distribution similar to the target population. Each record includes both name and ethnicity; the census sample file mentioned

above is a good example of such a file. The analyst ranks names by the strength of the association between name and ethnicity, e.g., almost everyone named 'Nguyen' is Vietnamese. All names with strength of association exceeding a chosen threshold and with frequency exceeding a chosen minimum are included on the list.

The Social Security Administration's file of applications for social security cards meets these criteria. It contains records for about 400 million social security number holders, alive and deceased. The file effectively is a registry of persons living in the United States since the inception of the social security program in 1936, but with significant undercoverage since some persons never applied for cards. The record content includes surname, maiden name, race in broad categories, and country of birth. Although ethnicity is not on the record, country of birth is a viable proxy for ethnicity for Asian Americans.

The data available for this project consisted of a subfile of applications by all persons born outside the United States before 1941 (originally extracted in 1995 to support actuarial estimates concerning the treatment of certain aliens under the social security program). We drew records from this subfile for all persons born in Asia and used this subfile to develop surname lists. The Asian subfile approximates the population of first-generation Asian Americans born before 1941, both alive and deceased. For women, we substituted maiden name for married surname.

A total of 1.8 million cardholders born before 1941 are native to one of the following 16 South and East Asian countries: Bangladesh, Burma, Cambodia, China (including People's Republic of China, Hong Kong and Taiwan), Indonesia, India, Japan, Korea (North and South), Laos, Malaysia, Pakistan, the Philippines, Singapore, Sri Lanka, Thailand, and Vietnam (North and South). The distribution by country of birth in Table 1 shows at least 130,000 records for each of the six countries of interest; together these six account for about 90 percent of the applicants born in Asia before 1941.

According to the 1990 census, the vast majority of Asian American elderly are foreign-born. Thus country of birth is a good proxy for ethnicity, the file of Asian-born persons includes a high proportion of Asian Americans born before 1941, and the ethnic distribution of persons in the file approximates the ethnic distribution of Asian American elderly in the general population. Japanese American elderly, however, are an exception since most are US-born. This exception could potentially bias our Japanese surname list derivation by an underestimation of the strength of association between Japanese country of birth and Japanese names. Fortuitously, Japanese names occur so infrequently among persons born in other Asian countries that we did not adjust for the under-representation of Japanese Americans in this file.

Table 1. Number of applicants for a social security card born before 1941 in Asia, by country of birth and sex

Place of birth	Males	Females	Total
Bangladesh	2,462	2,209	4,671
Burma	3,998	3,908	7,906
Cambodia	8,587	10,627	19,214
China	254,547	230,631	485,178
India	98,659	81,119	179,778
Indonesia	13,505	11,547	25,052
Japan	75,320	92,123	167,443
Korea	67,137	91,908	159,045
Laos	14,618	16,667	31,285
Malaysia	2,650	2,552	5,202
Pakistan	20,361	12,655	33,016
Philippines	237,263	250,557	487,820
Singapore	1,278	1,281	2,559
Sri Lanka	2,716	2,479	5,195
Thailand	9,277	12,366	21,647
Vietnam	62,358	68,057	130,415
Total	874,736	890,686	1,765,422

China includes Taiwan and Hong Kong. Korea includes North Korea and South Korea. Vietnam includes North Vietnam and South Vietnam.

We used the file of Asian-born cardholders to derive names for the context when race information is available. However, the derivation of name lists for use when no race identification is available required a file with racial and ethnic composition similar to the general population in the United States. Because the entire file of social security card applications was not available for this project, we turned to the Master Beneficiary Record (MBR), a file which includes persons entitled to social security benefits or enrolled in the Medicare program. Given the almost universal coverage by the Medicare program of those age 65 and older, we drew in October 1998 a subfile of over 70 million MBR records of persons born before 1934, ever enrolled in Part B of Medicare, and currently or (if deceased) last residing in the United States.

An MBR record includes surname and race – white, black or other – but not country of birth. To be of value for the derivation of name lists for Asian subgroups, a tabulation of the MBR by surname and race must be combined with the tabulation of surname and country of birth from the Asian-born file of cardholders. Our measure of the strength of association between a surname

and a specific Asian origin in a general population is the product A ∗ B, where A is the proportion with the associated Asian country of birth among persons with the specified surname in the file of Asian-born cardholders and B is the proportion with race 'other' among persons with that surname in the MBR. For example, in the file of Asian-born persons, 76 percent of persons with the surname 'Bang' are born in Korea, and in the MBR subfile, 22 percent of persons named 'Bang' have race code 'other'. Thus we estimate the proportion Korean of persons with the surname 'Bang' to be (0.76∗0.22), or 17 percent.

One complication to this strategy is that the 'other' race category includes not only Asian Americans, but also some Hispanic and Native American persons. The strategy would be compromised if Asian names also occurred among Hispanic and Native American persons. This is not a problem for Japanese, Chinese, Indian, Vietnamese and Korean names, but many Filipino names occur among Hispanic persons. Therefore, we took an additional step, removing names that appear on the 1990 Spanish surname list (Word & Perkins 1996) from the unconditional Filipino surname lists.

Before constructing the name lists, we eliminated any name that occurred fewer than five times in the file of Asian-born persons. Then for both the lists conditional on Asian race and the lists not conditional on race, a surname was included if *at least* 50 percent of persons with that surname were associated with an origin (e.g., Korea) and *less than* 50 percent with any of the other countries. These lists we call 'predictive'. A subset of names from each list was further identified as 'strongly predictive' by using a threshold of 75 percent. A few surnames selected for conditional lists did not appear in the MBR subfile; we included such names in the predictive unconditional lists only when they were in the strongly predictive conditional list.

We developed 24 lists in all: two sets (predictive and strongly predictive) of two types (conditional and unconditional) for six Asian American groups. The progression from predictive to strongly predictive improves accuracy, but at the cost of reduced coverage. Thus the two sets of lists are suited to different applications, dictated by the importance of accuracy (e.g., being surer of a person's Chinese ethnicity versus detecting a higher proportion of Chinese persons).

Evaluation of surname lists

We evaluated the 24 lists with regard to sensitivity (coverage) and positive predictive value (accuracy). The sensitivity measure for a list is the proportion of all persons of the given origin whose name appears on the list. The positive predictive value (PPV) is the proportion of persons with names on the list who are of that origin (Figure 1). Recall that these measures necessarily refer to country of birth as the proxy for ethnicity (a limitation inherent in the SSA

	Target birthplace	Other birthplaces
Specified surnames	a	b
Other surnames	c	d

Figure 1. Sensitivity and positive predictive value of surname lists. Sensitivity = $a/(a + c)$; Positive predictive value = $a/(a + b)$.

source file). As an independent check, we turned next to the 1.9 million record census file used to derive the 1990 Spanish surname list. Although too small for the derivation of Asian name lists, this file is ample in size for evaluation. The file had already been tabulated to obtain for each surname the total number of persons, the number who identified themselves as belonging to one of the six Asian ethnic groups under study, and the number who identified themselves as belonging to other Asian groups. Because the census sample consists of a cross-section of adults, it affords the opportunity to evaluate list performance in a population that includes non-elderly as well as elderly and native-born as well as foreign-born Asian Americans.

For comparison, we also evaluated with the census file two previously published Chinese surname lists and some preliminary lists developed at the Census Bureau in the 1980s. (Note that we did not have access to the census file; Census Bureau staff graciously calculated the summary measures needed for the lists we furnished.)

Results

Conditional lists

After eliminating those surnames which occurred fewer than five times, about 27,000 surnames remained, which accounted for 86 percent of the 1.8 million older social security cardholders born in Asia. Only six surnames were extremely common, being held by more than 10,000 persons each: Chan, Chang, Chen, Lee, Nguyen, Wong. Almost 21,000 of the 27,000 surnames were predictive of a single Asian country of birth and were therefore included on one of the six predictive conditional lists. Of 168 names occurring more than 1,000 times each, six (Ha, Jung, Ko, Lee, Lim, Tan) were not predictive, owing to their distribution across several Asian countries.

The six predictive, conditional lists vary dramatically in length (Table 2). The lists for Korean and Vietnamese origins consist of fewer than 400 names, while the list for Filipino origin contains more than 12,000 names. (The 50 most common names on each of these lists are given in the Appendix.)

Table 2. Number of surnames on the 24 predictive and strongly predictive lists, and their sensitivity and positive predictive value in the source data file

	Predictive			Strongly predictive		
Country of Birth	Number of names	SE	PPV	Number of names	SE	PPV
Conditional on birth in Asia						
China	1200	0.78	0.89	902	0.69	0.93
India	2797	0.60	0.83	2198	0.48	0.91
Japan	3559	0.79	0.96	3465	0.79	0.96
Korea	288	0.64	0.82	205	0.50	0.90
Philippines	12475	0.76	0.98	12314	0.75	0.99
Vietnam	374	0.79	0.86	231	0.68	0.91
Unconditional						
China	791	0.72	0.81	461	0.57	0.88
India	2051	0.43	0.74	977	0.24	0.87
Japan	3369	0.77	0.89	2634	0.71	0.92
Korea	209	0.52	0.74	110	0.36	0.83
Philippines	8654	0.32	0.83	6649	0.25	0.91
Vietnam	249	0.74	0.84	95	0.61	0.89

Sensitivity (SE) is the proportion of all persons of a given country of birth whose names appear on the list. It is a measure of coverage. Positive predictive value (PPV) is the proportion of persons whose names appear on the list who were born in the corresponding country. It is a measure of accuracy.

The PPV for the six predictive lists (a summary measure of their accuracy) varies from a high of 98 percent for the Filipino to 82 percent for the Korean, with an average of 89 percent. The PPV is 90 percent or more for all of the strongly predictive sublists.

The sensitivity of the predictive lists (their overall completeness) for Chinese, Filipino, Japanese, and Vietnamese origins is between 75 and 80 percent, but is lower for Indian and Korean origins, 60 and 64 percent. Incomplete coverage may be attributed to one of two circumstances: surnames that are rare (omitted due to the minimum occurrence threshold) or surnames that are not strongly associated with a single origin.

Interestingly, Japanese and Filipino surnames are so distinctive among the Asian-born that nearly all of the names on the predictive lists are also on the strongly predictive sublists. For the other four origins sensitivity decreases noticeably in progressing from the predictive to the strongly predictive lists.

Nevertheless, each of the strongly predictive lists registers at least 48 percent sensitivity.

Unconditional lists

We also constructed lists for use when no race information is available. Here sensitivity was at least 50 percent for four of the six predictive lists and three of the six strongly predictive lists (lower panel of Table 2). The lower sensitivity of the unconditional lists is not surprising since the unconditional lists were derived by eliminating names not distinctly Asian. The number of names not distinctly Asian is relatively few for Japanese, Vietnamese, and Chinese origins. The number is substantial for the Filipino origin, because of overlap between Filipino and Spanish surnames. The PPV of each of the unconditional lists is also somewhat lower than for the corresponding conditional list; it averages 81 percent.

Evaluation with census records

Tables 3 and 4 show corresponding summary performance measures derived from the census sample, which affords an independent basis for evaluation. Table 3 refers to a conditional situation, that is, the application of the conditional lists to a subset of persons who self-identified as belonging to any Asian subgroup on the 1990 census. Table 4 refers to the unconditional situation – the application of the unconditional lists to the entire census sample, both Asian and non-Asian. As previously described, the census sample covers all householder ages, includes native-born Asian Americans, and has explicit rather than proxy ethnicity information. Thus it is encouraging to note how similar the coverage and accuracy measures are to those derived from the Social Security Administration source files.

For the conditional lists, the PPVs tend to be higher, but the sensitivity is somewhat lower for the census file than for the source file. The PPVs average 92 percent for the predictive lists and 95 percent for the strongly predictive lists. The sensitivity averages just 3 percentage points less for the census file than for the source file across the 12 conditional lists.

The coverage and accuracy measures of the unconditional lists in the census file are also markedly similar to those derived from the source files. Across the 12 unconditional lists, the average sensitivity was less than one percentage point lower in the census file; the average PPV was about the same.

As a further step, for comparison, we also evaluated several other unconditional lists previously derived for English-speaking areas. First, we considered two lists of Chinese surnames: a list Hage and others developed in

Table 3. Sensitivity and positive predictive value of the 12 conditional lists in a sample of 1990 census records, among all householders who identified themselves as belonging to any Asian subgroup

	Predictive		Strongly predictive	
Census race subgroup	SE	PPV	SE	PPV
Chinese	0.74	0.89	0.65	0.93
Indian	0.51	0.98	0.42	0.99
Japanese	0.73	0.99	0.73	0.98
Korean	0.63	0.87	0.54	0.93
Filipino	0.71	0.93	0.70	0.94
Vietnamese	0.87	0.87	0.70	0.91

SE – sensitivity; PPV – positive predictive value.

Table 4. Sensitivity and positive predictive value of the 12 unconditional lists in a sample of 1990 census records among all householders

	Predictive		Strongly predictive	
Census race subgroup	SE	PPV	SE	PPV
Chinese	0.70	0.76	0.55	0.83
Indian	0.38	0.77	0.24	0.82
Japanese	0.71	0.92	0.68	0.93
Korean	0.54	0.81	0.43	0.88
Filipino	0.29	0.86	0.24	0.84
Vietnamese	0.74	0.83	0.65	0.88

SE – sensitivity; PPV – positive predictive value.

Australia, not by an 'empirical' method, but by determining English spellings for different dialectic pronunciations of 33 common Chinese-character family names and then supplementing with other names from the telephone directory (Hage et al. 1990); and a list derived by Choi and others in Canada from an Ontario mortality database containing surname and country of birth for about 4,000 decedents born in China (China, Taiwan or Hong Kong). Names were ranked by positive likelihood ratio (the ratio of the post-test odds to the pre-test odds) (Choi et al. 1993).

The Hage list had a sensitivity of 59 percent and a PPV of 44 percent when applied to the census database (Table 5). Dividing the Choi list into three levels by positive likelihood ratio, the sensitivities ranged from 62 percent for

Table 5. Sensitivity and positive predictive value of previously derived lists in a sample of 1990 census records, among all householders

Surname list	SE	PPV
Hage Chinese List	0.59	0.44
Choi Chinese List 1	0.62	0.64
Choi Chinese List 2	0.69	0.58
Choi Chinese List 3	0.78	0.28
1982 Seed Lists		
Chinese American	0.45	0.79
Indian American	0.26	0.78
Japanese American	0.41	0.93
Korean American	0.56	0.69
Filipino American	0.08	0.61
Filipino American (excluding Spanish names)	0.05	0.76
Vietnamese American	0.70	0.86

SE – sensitivity; PPV – positive predictive value. The Hage list (Hage et al. 1990) includes all romanized Chinese names in their Table 1. Choi lists (Choi et al. 1993) are derived as follows from their Table 2 (males) and Table 3 (females): List 1 includes names with cutoff positive likelihood ratio of 400 or higher on either Table 2 or 3; List 2 includes names with cutoff positive likelihood ratio of 100 or higher on either Table 2 or 3; and List 3 includes all listed names on Table 2 or 3. The seed lists were presented at the 1982 meeting of the Population Association of America (Passel et al. 1982).

the least comprehensive to 78 percent for the most comprehensive list. The corresponding PPVs ranged from 64 percent down to 28 percent for the most comprehensive list.

Next we considered the results of a noteworthy pilot project by Passel and others at the Census Bureau in the early 1980s aimed at eventually building Asian surname lists (Passel et al. 1982). They developed preliminary or 'seed' lists of surnames for twelve Asian groups. The lists were compiled from the 1979 Alien Address File, which included country of origin; these lists were intended as starting points to derive geographic profiles for each group, which would then be used to augment and refine the lists. Though never published,

the seed lists have been used in at least two research projects (Lauderdale et al. 1997; Rosenwaike 1994). The seed lists are much shorter than our lists and are largely subsets of them. The Korean and Vietnamese seed lists have a sensitivity similar to that of our lists when applied to the census dataset; the other four groups have a lower sensitivity than our lists (Table 5). Accuracy of the seed lists is lower for Korean and Filipino surnames.

Discussion

We have derived and separately evaluated surname lists for identifying persons belonging to the six principal Asian American ethnic groups: Chinese, Japanese, Asian Indian, Korean, Filipino and Vietnamese. These lists permit identification both in contexts where race information is available and in those where it is not. An evaluation against an independent census benchmark (of all Asian Americans) demonstrates that the conditional lists identify, with high accuracy, a majority of persons who self-identify as belonging to each ethnic group. In populations containing Asians as well as non-Asians, the majority of Chinese, Korean, Japanese, and Vietnamese can still be identified, as well as substantial proportions of Filipino and Asian Indians.

Race information increases the potential of surnames for ethnic identification. Previous work has not explicitly recognized the advantage of using race information when available. This advantage might be extended not only to individual records with race information but also to records of persons residing in census tracts or zip codes with a high proportion of Asian Americans. Both coverage and accuracy are greater if Asian race information is available, particularly for Filipino Americans.

The six surname lists, together containing 20,693 unique names, are available from the first author as an electronic text file. Although these six groups represent about 90 percent of Asian Americans, we do not recommend combining the lists to identify an aggregated group of Asian Americans without adjusting for differences in the sensitivity of the six lists. Otherwise, ethnicities for whom surname identification has greater sensitivity would be over-represented in the aggregate.

We evaluated our lists, and previously-derived ones, with tabulations of census records augmented with surname in conjunction with the 1990 post-enumeration survey. Our list of Chinese surnames identified a larger proportion of the ethnic sample with higher accuracy than either of the two previously published lists. The unpublished census seed lists are comparable to our lists for Korean and Vietnamese names, while each of our other four lists had greater coverage than the corresponding seed list.

Perkins (1993) used the same sample of 1990 census records to evaluate the 1980 Spanish surname list. His finding of 80 percent sensitivity and 90 percent accuracy provides a yardstick for appreciating these values for our lists. Our unconditional Chinese, Japanese, and Vietnamese predictive lists have accuracy similar to and coverage somewhat lower than the Spanish list, while the coverage for the other three lists is considerably poorer. However, when race data are available, the predictive conditional lists for Chinese, Japanese, Filipino, and Vietnamese are as good or better than the Spanish list in coverage and accuracy.

Without race information, the Asian lists do not identify as high a percentage of the population as does the Spanish surname list. The primary reason is that many frequently-occurring Asian surnames either are not unique to a single Asian ethnic group or are not uniquely Asian. The overlap between Hispanic and Filipino surnames poses a much greater problem for identifying Filipinos than for identifying Hispanics because of the relative sizes of the populations. Omitting non-specific names limits coverage. A second reason is that because of the relatively small size of these populations many names fail to meet a minimum occurrence threshold. Third, country of birth is an imperfect proxy for ethnicity. For example, large numbers of ethnic Chinese were born in the Philippines, Singapore, Malaysia and Vietnam; and ethnic Koreans were born in Japan, China, and Russia. Furthermore, a disproportionate number of Asian immigrants to the United States may come from minority Asian populations residing in different Asian countries.

The primary limitation of surname identification arises from marriage outside the ethnic group and concomitant name change for women. We cannot quantify the difference in accuracy or coverage for women relative to men that results from out-marriage with the census evaluation file because the file was limited to heads of household, unrelated individuals and unmarried females. Some data sources include maiden name, whose use would decrease the difference in accuracy and coverage. However, the difference should be relatively small when the study population is elderly or predominantly foreign-born. According to a tabulation of the 1990 census public use file, among elderly married men in these six ethnic groups, the proportion married to a woman of the same ethnic group ranges from 96 percent of Chinese to 88 percent of Filipino. For elderly married women, the proportions of same-ethnicity marriage range from 94 percent of Chinese to 87 percent of Japanese and 84 percent of Korean (Teresa Labov and Jerry A. Jacobs, University of Pennsylvania, personal communication, 1997). The higher rates of out-marriage for elderly Japanese and Korean women may be due to the war brides phenomenon. Across the entire age range, 80 percent or more of married persons in each sex-ethnicity group are married to a person of the same

ethnicity for all groups except Filipino women (63%), Japanese men (75%), and Japanese women (55%). For Japanese, the high rates of out-marriage are likely related to the high proportion US-born. While only 13 percent of first generation Asian Americans marry non-Asians, the percentage increases with generation (Smith & Edmonston 1997: 132).

A separate question is what proportion of women who have out-married retain their maiden name; we are unaware of any data concerning this. Name change at marriage is not the norm in some Asian countries (Rutledge 1992: 153).

One question that arises from the application of these lists is whether persons with surnames on the list might differ from persons of the same ethnicity with surnames not on the list. For example: do the 70 percent of Chinese Americans with distinctly Chinese surnames differ from the 30 percent with non-distinctive names (such as Chang or Lee) or uncommon names? There is little research addressing this issue. Shin & Yu (1984) demonstrated that Koreans with the single surname 'Kim' formed like proportions (20 to 23 percent) among Korean directories defined by occupation, residence and institutional affiliation, suggesting that the distribution of this very common Korean surname did not differ across social strata. For Asian Indians, though, names derived from Arabic (and held by Muslims) are probably less likely to be distinctively Indian than names derived from indigenous Indian languages (and held by Hindus). A greater concern is that Asian women married to Asian men are likely to differ in many ways from women married to non-Asians.

The surname lists developed and evaluated here will offer researchers a variety of applications, which generally fall into three categories: identification of individuals to target for study participation; inference of ethnicity in data sources lacking ethnic detail; and characterization of the ethnic composition of a population. First, the lists could be used to find persons likely to belong to a particular ethnic group from sources such as telephone listings or clinic records. These individuals would subsequently be contacted to corroborate ethnicity before enrollment in a survey or panel. The lists could be a means to create an ethnically homogenous study population or to oversample by ethnicity. Second, the lists could be used to infer ethnicity when direct corroboration is not feasible, for example from death certificates, disease registries or other administrative data. One could thus contrast morbidity or mortality rates for ethnic groups from administrative data lacking ethnic information. Finally, the lists could be used (possibly in conjunction with the sensitivity and PPV from the census evaluation) with rosters of students, voters, homeowners or drivers in an area to estimate the size of an ethnic pop-

ulation, to compare the relative concentrations of ethnic groups in different areas, or to track the size of an ethnic population over time.

Acknowledgments

The research of the first author was supported by the National Institute on Aging (Grant R03-AG14871-01). The opinions expressed in this article are those of the authors, and no official endorsement by the Social Security Administration should be inferred. The authors are indebted to Barry Bye and the Bureau of the Census for tabulating census records. The authors thank Ira Rosenwaike, Dr Jack Goldberg and Dr Kate Cagney for their comments on the project. The authors also thank an anonymous reviewer for editorial suggestions. Parts of this paper were presented at the 1999 annual meetings of the Population Association of America in New York City and the American Public Health Association in Chicago.

Remark. The complete surname lists are available from the corresponding author via electronic mail: <lauderdale@health.bsd.uchicago.edu>.

Appendix

50 most frequently-occurring names in each conditional surname list.

Chinese

1. Wong	11. Ng	21. Lau	31. Chiu	41. Yuen
2. Chen	12. Yu	22. Fong	32. Lai	42. Chao
3. Chan	13. Cheng	23. Leung	33. Tam	43. Kwan
4. Wang	14. Yee	24. Chow	34. Lo	44. Tong
5. Chang	15. Yang	25. Cheung	35. Tsai	45. Shen
6. Lin	16. Chu	26. Tang	36. Liang	46. Kuo
7. Wu	17. Chin	27. Lu	37. Woo	47. Louie
8. Liu	18. Ho	28. Sun	38. Chou	48. Moy
9. Huang	19. Lam	29. Ma	39. Hu	49. Eng
10. Li	20. Hsu	30. Zhang	40. Chiang	50. Kwong

Japanese

1. Suzuki	11. Yamada	21. Abe	31. Murakami	41. Harada
2. Sato	12. Yoshida	22. Ikeda	32. Ishii	42. Takeuchi
3. Tanaka	13. Kato	23. Inoue	33. Yamashita	43. Fujii
4. Takahashi	14. Kimura	24. Hashimoto	34. Nishimura	44. Aoki
5. Watanabe	15. Matsumoto	25. Ogawa	35. Kondo	45. Matsuda
6. Nakamura	16. Hayashi	26. Ono	36. Fujita	46. Okamoto
7. Yamamoto	17. Sasaki	27. Ishikawa	37. Nakagawa	47. Goto
8. Kobayashi	18. Yamaguchi	28. Okada	38. Sakai	48. Tamura
9. Ito	19. Mori	29. Sakamoto	39. Nakajima	49. Arai
10. Saito	20. Shimizu	30. Maeda	40. Hasegawa	50. Takeda

Korean

1. Park	11. Song	21. Yun	31. Rhee	41. Jang
2. Kim	12. Shin	22. Suh	32. Won	42. Hyun
3. Choi	13. Oh	23. Son	33. Yim	43. Whang
4. Cho	14. Yoon	24. An	34. Kwak	44. Huh
5. Chung	15. Hwang	25. Cha	35. Shim	45. Chae
6. Kang	16. Yoo	26. Min	36. Jun	46. Mun
7. Yi	17. Choe	27. Nam	37. Sin	47. No
8. Han	18. Kwon	28. Bae	38. Paik	48. Sim
9. Pak	19. Ahn	29. Im	39. Seo	49. Sohn
10. Hong	20. Chun	30. Chon	40. Bang	50. O

Indian

1. Singh	11. Ahmed	21. Chacko	31. Trivedi	41. Vyas
2. Shah	12. Parikh	22. Dave	32. Das	42. Fernandes
3. Khan	13. Hussain	23. Varghese	33. Pandya	43. Grewal
4. Patel	14. Joshi	24. Sheth	34. Sandhu	44. Qureshi
5. Ali	15. Amin	25. Jain	35. Iyer	45. Chand
6. Desai	16. Bhatt	26. Lal	36. Siddiqui	46. Dhillon
7. Mehta	17. Gandhi	27. Mathai	37. Kumar	47. Ullah
8. Rao	18. Ram	28. Husain	38. Parekh	48. Mistry
9. Sharma	19. Ahmad	29. Bhakta	39. Sidhu	49. Nair
10. Gupta	20. Mathew	30. John	40. Prasad	50. Hasan

Filipino

1. Reyes	11. Flores	21. Castro	31. Hernandez	41. Diaz
2. Santos	12. Gonzales	22. Santiago	32. Valdez	42. Pascua
3. Garcia	13. Villanueva	23. Tolentino	33. Pascual	43. Gutierrez
4. Cruz	14. Lopez	24. Delrosario	34. Ramirez	44. Velasco
5. Ramos	15. Deleon	25. Torres	35. Francisco	45. Antonio
6. Delacruz	16. Castillo	26. Soriano	36. Corpuz	46. Angeles
7. Mendoza	17. Aquino	27. Sanchez	37. Mercado	47. Morales
8. Bautista	18. Rivera	28. Martinez	38. Navarro	48. Dejesus
9. Deguzman	19. Domingo	29. Rodriguez	39. Javier	49. Manuel
10. Fernandez	20. Perez	30. Dizon	40. Ocampo	50. Mariano

Vietnamese

1. Nguyen	11. Dang	21. Doan	31. Diep	41. Phu
2. Tran	12. Do	22. Dao	32. Ton	42. Vinh
3. Le	13. Bui	23. Thai	33. La	43. Quang
4. Pham	14. Vo	24. Mai	34. Thach	44. Tieu
5. Huynh	15. Ly	25. Van	35. Thi	45. Hoa
6. Vu	16. Duong	26. Cao	36. Thanh	46. Trang
7. Phan	17. Luong	27. Vuong	37. Dam	47. Giang
8. Truong	18. Dinh	28. Phung	38. Vong	48. Luc
9. Hoang	19. Trinh	29. Quach	39. Trieu	49. Banh
10. Ngo	20. Luu	30. Ta	40. Buu	50. Nghiem

References

Abrahamse A. F., Morrison, P. A. & Bolton, N. M. (1994). Surname analysis for estimating local concentration of Hispanics and Asians, *Population Research and Policy Review* 13: 383–398.

Barringer, H. R., Gardner, R. W. & Levin, M.J. (1993). *Asians and Pacific Islanders in the United States. The Population of the United States in the 1980s*. New York: Russell Sage Foundation.

Buechley, R. W. (1976). Generally useful ethnic search system: GUESS. University of New Mexico, Cancer Research and Treatment Center (mimeo).

Choi, B. C. K., Hanley, A. J., Holowaty, E. J. & Dale, D. (1993). Use of surnames to identify individuals of Chinese ancestry, *American Journal of Epidemiology* 138: 723–734.

Hage, B. H. H., Oliver, R. G., Powles, J. W. & Wahlqvist, M. L. (1990). Telephone directory listings of presumptive Chinese surnames: an appropriate sampling frame for a dispersed population with characteristic surnames, *Epidemiology* 1: 405–408.

Kuo, J. & Porter, K. (1998). Health status of Asian Americans: United States, 1992–1994. Hyattsville, MD: National Center for Health Statistics, Advance data from vital and health statistics, No. 298.

Lauderdale, D. S., Jacobsen, S. J., Furner, S. E., Levy, P. S., Brody, J. A. & Goldberg, J. (1997). Hip fracture incidence among elderly Asian American populations, *American Journal of Epidemiology* 146: 502–509.

Passel, J. S. & Word, D. L. (1980). Constructing the list of Spanish surnames for the 1980 Census: an application of Bayes theorem. Paper presented at the annual meeting of the Population Association of America, Denver, Colorado, April 1980.

Passel, J. S., Word, D. L., McKenney, N. D. & Kim, Y. (1982). Postcensal estimates of the Asian population in the United States: description of methods using surname and administrative records. Paper presented at the annual meeting of the Population Association of America, San Diego, California, April 1982.

Perkins, R. C. (1993). Evaluating the Passel-Word Spanish surname list: 1990 decennial census post enumeration survey results. Washington, DC: US Bureau of the Census, Population Division, Technical Working Paper No. 4.

Rosenwaike, I. (1994). Surname analysis as a means of estimating minority elderly: an application using Asian surnames, *Research on Aging* 16: 212–227.

Rutledge, P. J. (1992). *The Vietnamese experience in America*. Bloomington and Indianapolis, IN: Indiana University Press.

Shin, E. H. & Yu, E. Y. (1984). Use of surnames in ethnic research: the case of Kims in the Korean-American population, *Demography* 21: 347–359.

Smith, J. P. & Edmonston, B., eds. (1997). *The new Americans: economic, demographic, and fiscal effects of immigration.* Washington, DC: National Academy Press.

US Department of Health and Human Services, Public Health Service (1991). Healthy people 2000: national health promotion and disease prevention objectives. Washington, DC: Government Printing Office.

Word, D. L. & Perkins Jr., R.C. (1996). Building a Spanish surname list for the 1990s: a new approach to an old problem. Washington, DC: US Census Bureau, Technical Working Paper 13.

Yu, E. S. H. & Liu, W. T. (1992). US national health data on Asian Americans and Pacific Islanders: a research agenda for the 1990s, *Amercan Journal of Public Health* 82: 1645–1652.

Address for correspondence: Diane S. Lauderdale, Department of Health Studies, University of Chicago, 5841 S. Maryland Ave., MC 2007, Chicago, IL 60637, USA
Phone: (773) 834-0913; Fax: (773) 702-1979;
E-mail: lauderdale @ health.bsd.uchicago.edu