

NAACCR Asian/Pacific Islander Identification Algorithm [NAPIIA v1.2.1]: Enhancing the Specificity of Identification

August 12, 2011



Editors:

NAACCR Race and Ethnicity Work Group

Chaired by
Francis P. Boscoe, PhD
New York State Cancer Registry

Suggested Citation:

NAACCR Race and Ethnicity Work Group. *NAACCR Asian Pacific Islander Identification Algorithm [NAPIIA v1.2.1]*. Springfield, IL: North American Association of Central Cancer Registries, August 2011.

Cooperative Agreement Number U75/CCU523346 from CDC provided funds to NAACCR for statistical support for development of the algorithm. The contents of the report are solely the responsibility of the authors and do not necessarily represent the official views of CDC.

The NAACCR Race and Ethnicity Work Group

Catherine S. Grafel-Anderson
Hawaii Tumor Registry
cganderson@crch.hawaii.edu

Peg Balcius
Los Angeles Cancer Surveillance Program
balcius@usc.edu

Francis P. Boscoe, PhD
NY State Cancer Registry (chair)
fpb01@health.state.ny.us

Michael Green
Hawaii Tumor Registry
Michael@crch.hawaii.edu

Mei-chin Hsieh, MSPH
Louisiana Tumor Registry
mhsieh@lsuhsc.edu

Andrew Lake
IMS, Inc.
lakea@imsweb.com

George J. Lara, MA
Texas Cancer Registry
George.Lara@dshs.state.tx.us

Lihua Liu, PhD
Los Angeles Cancer Surveillance Program
lihualiu@usc.edu

Barry Miller, DrPH
SEER Program
millerb@mail.nih.gov

Paulo Pinheiro, MD, PhD
Florida Cancer Data System
ppinheiro@med.miami.edu

Maria J. Schymura, PhD
NY State Cancer Registry
mjs08@health.state.ny.us

Sarah Shema
Northern California Cancer Center
sshema@nccc.org

Cheryll Thomas, MSPH
Centers for Disease Control and Prevention
zzg3@cdc.gov

Contributors to previous editions

Vivien W. Chen
Holly L. Howe
Betsy Kohler
Arti Parikh

NAACCR Asian Pacific Islander Identification Algorithm (NAPIIA) v1.2.1

Summary

The NAACCR Asian Pacific Islander Identification Algorithm version 1 (NAPIIA v1.2.1) uses a combination of NAACCR variables to classify cases directly or indirectly as Asian/Pacific Islander for analytic purposes. It is focused on coding cases with a race code of Asian NOS (race code 96) or Pacific Islander NOS (race code 97) to a more specific Asian or Pacific Islander race category, using the birthplace and name fields (first, last, and maiden names). Birthplace can be used to indirectly assign a specific race to one of eight Asian groups (Chinese, Japanese, Vietnamese, Korean, Asian Indian, Filipino, Thai, and Cambodian) and three Pacific Islander groups (Samoan, Micronesian, and Polynesian). Names can be used to indirectly assign a specific race to one of seven Asian groups (Chinese, Japanese, Vietnamese, Korean, Asian Indian, Filipino, and Hmong) and three Pacific Islander groups (Hawaiian, Guamanian, and Samoan). The next version of NAPIIA (2.0), slated for release in 2011, will allow the recoding of cases coded as 98 (Other) or 99 (Unknown) to 96 or 97 based on name and birthplace.

The algorithm uses the following NAACCR standard variables:

- Race 1 through Race 5 (Items 160 through 164)
- Spanish/Hispanic Origin (Item 190)
- Name – Last (Item 2230)
- Name – First (Item 2240)
- Name – Maiden (Item 2390)
- Birthplace (Item 250)
- Sex (Item 220)

What's new in version 1.2.1

This version of NAPIIA contains **three** enhancements over the previous version (1.2):

1. The algorithm is now compatible with both NAACCR Record Layout Versions 11.3 and 12.
2. For cases diagnosed on or after January 1, 2010, code 09 (Asian Indian/Pakistani) has been retired and replaced with the following new codes: Asian Indian/Pakistani NOS (15), Asian Indian (16), and Pakistani (17). Birthplace rules have also been updated to reflect this change.
3. Cases with Race 1 equal to 96 (Asian, NOS) or 97 (Pacific Islander, NOS) also coded as Hispanic will retain the 96 or 97 code. Previously, these cases would typically be recoded as Filipino based on surname. Several researchers indicated that this is not accurate. This change also simplifies the relationship between NAPIIA and NHIA.

This version of the documentation (August 2011) adds item number 3 above, which was omitted from the previous edition of the documentation (September 2010). The algorithm itself has not changed between 2010 and 2011.

What's coming in future versions

1. Application of the algorithm to cases coded as 98 (other race) and 99 (unknown). This is being evaluated at the time of this writing.
2. Conversion from registry-specific country codes to country codes following the ISO 3166 standard.

Detailed NAPIIA v1.2.1 Logic

Step 1. Identify cases containing race code 96 or 97

1.1. Single race code of 96. All cases with a Race 1 code (data item 160) of 96 and no additional race codes **and not identified as Hispanic (data item 190 equal to 0, 7 or 9)** will be identified and retained for Steps 3 and 4 of the algorithm. For these cases, the codes for Race 2 through Race 5 (data items 161-164) must be blank or 88.

1.2. Single race code of 97. All cases with a Race 1 code (data item 160) of 97 and no additional race codes **and not identified as Hispanic (data item 190 equal to 0, 7 or 9)** will be identified and retained for Steps 3 and 4 of the algorithm. For these cases, the codes for Race 2 through Race 5 (data items 161-164) must be blank or 88.

1.3 Race code of 96 or 97 in combination with one or more other race codes

Evaluated in this step are records that have at least two of the five race data items (items 160 through 164) filled with values other than blank or 88, at least one of which is coded with 96 and/or 97. The various scenarios are presented in Table 1.3. For some rare and unusual scenarios, Race 1 (item 160) is given precedence, but these cases should also be reviewed manually, since a coding error may be likely. In the event that cases are revised as a result of manual review, a new data set should be created and the NAPIIA algorithm should be restarted. For further guidance on race coding issues, consult the SEER Program Coding and Staging Manual, 2004, pp. 46-50¹. Note that as of 2009, very few cancer cases are reported to central registries with more than one race, with the number of cases in the entire US well below 0.5% of total cases reported. If reporting of multi-race cases becomes more common in the future, this step will be re-evaluated for continued appropriateness and validity.

Scenario	Action
1.3.1. One race code is 04-32, one race code is 96 or 97; others are blank or 88.	The 04-32 takes precedence. Treat as a single race case. Go to step 2.
1.3.2. One race code is 07; one race code is 96 or 97; others are any value.	The 07 code takes precedence.
1.3.3. More than one race code is 04-32 excepting 07; one race code is 96 or 97; others are blank or 88.	The contents of Race 1 are output, and the case is flagged for manual review.
1.3.4. One race code is 01; one race code is 96 or 97; others are blank or 88.	The 96 or 97 code takes precedence. Go to step 3.
1.3.5. One or more race codes is 02-03; one race code is 96 or 97; others are blank or 88.	The contents of Race 1 are output, and the case is flagged for manual review.
1.3.6. One race code is 96; one race code is 97; others are any value.	The contents of Race 1 are output, and the case is flagged for manual review.
1.3.7. Any multiple race combination involving code 96 or 97 not listed above.	The contents of Race 1 are output, and the case is flagged for manual review.

Step 2. Directly code cases not containing code 96 or 97.

2.1 Direct Code Single Race Cases

Directly code all single race cases in this step. These consist of all cases with codes 01 through 32, 98, and 99 in Race 1 (data item 160), **and cases with codes 96 or 97 in Race 1 identified as Hispanic (data item 190 equal to 1-6 or 8)** (Table 2.1). For these cases, the Race 2 through Race 5 fields (data items 161-164) must be blank or 88, unless Race 1 is 99, in which case Race 2 through Race 5 should also be 99.

Code	Category
01	White
02	Black
03	American Indian, Aleutian, or Eskimo (includes all indigenous populations of the Western hemisphere)
04	Chinese
05	Japanese
06	Filipino
07	Hawaiian
08	Korean
10	Vietnamese
11	Laotian
12	Hmong
13	Kampuchean
14	Thai
15	Asian Indian or Pakistani (was previously code 09)
16	Asian Indian
17	Pakistani
20	Micronesian, NOS
21	Chamorroan
22	Guamanian, NOS
25	Polynesian, NOS
26	Tahitian
27	Samoan
28	Tongan
30	Melanesian, NOS
31	Fiji Islander
32	New Guinean
96	Asian, NOS (Hispanic only)
97	Pacific Islander, NOS (Hispanic only)
98	Other
99	Unknown

2.2 Multiple Race Cases

Evaluated in this step are records that have at least two of the five race data items (items 160 through 164) filled with values other than blank or 88. Refer to Table 2.2. Again, if reporting of multi-race cases becomes more common in the future, then this step will be re-evaluated for continued appropriateness and validity.

Scenario	Action
2.2.1. One race code is 01, one race code is 02-32; others are blank or 88.	The 02-32 takes precedence.
2.2.2. One race code is 07, others may be any value.	The 07 takes precedence.
2.2.3. All other multiple race combinations.	The contents of Race 1 are output, and the case is flagged for manual review.

Step 3. Indirect Identification Based on Birthplace

The indirect identification component of NAPIIA v1.2 is applied only to single race persons that have a code of 96 or 97 on NAACCR Standard data item 160 as identified in Steps 1.1-1.2, or certain multiple race persons identified in Step 1.3.

3.1 Included Asian and Pacific Island Birthplaces

If a person has a race and birthplace combination of any of the countries listed in Table 3.1, the person should be coded to the specific Asian or Pacific Islander race group as designated in the table^a. These persons have a high probability of being the specific Asian or Pacific Islander race group listed.

Existing Race	Birthplace	Code	Recoded Race
96	China, Taiwan, Hong Kong, Macao	681, 682, 683, 684, 686	Chinese
96	Nampo-Shoto, Ryukyu Islands, Japan	133, 134, 693	Japanese
96	Philippines	675	Filipino
96	Korea, North Korean, South Korean	695	Korean
96	Pakistan	639	Pakistani
96	India	641	Asian Indian
96	Vietnam	665	Vietnamese
96	Thailand	651	Thai
96	Cambodia, Kampuchea		Cambodian or Kampuchean
97	American Samoa	121	Samoan
97	Kiribati, Micronesia, Johnson Atoll, Marshall Islands, Palau, Micronesian Islands	122, 123, 127, 131, 139, 723	Micronesian, NOS
97	Cook Islands, Tuvalu, Tokelau, Polynesian Islands	124, 125, 136, 725	Polynesian, NOS

3.2. Excluded Asian and Pacific Island Birthplaces

If a person has a birthplace (data item 250) that is considered non-predictive, race code 96 should be retained and no further steps in the algorithm performed (Table 3.2). These birthplaces are too ambiguous or suggest race groups for which no code exists (e.g., Malay).

Code	Birthplace
640	Maldives
643	Nepal, Bhutan
645	Bangladesh
647	Sri Lanka
649	Myanmar/Burma
671	Malaysia, Singapore, Brunei
673	Indonesia
685	Tibet
691	Mongolia

A name can still be predictive even when the birthplace is not predictive. For example, a person with the surname Chang born in Malaysia is highly likely to be Chinese. However, to be conservative and consistent with the SEER Coding and Staging Manual, such cases are not recoded in NAPIIA version 1.2.1.^b

3.3. Excluded Hispanic Birthplaces

Cases with a birthplace that is highly predictive of Hispanic ethnicity are also excluded from indirect identification. While rare, such cases would be highly likely to be recoded to Filipino based on their name. Considering that Filipino migration to Latin America has been historically negligible, the 96 code for such cases is uninformative. Table 3.3 lists these birthplaces.

101	Puerto Rico
230	Mexico
241	Cuba
243	Dominican Republic
250	Central America
251	Guatemala
253	Honduras
254	El Salvador
255	Nicaragua
256	Costa Rica
257	Panama
265	Latin America, NOS
300	South America
311	Colombia

321	Venezuela
345	Ecuador
351	Peru
355	Bolivia
361	Chile
365	Argentina
371	Paraguay
375	Uruguay
443	Spain, Andorra

Step 4. Indirect Identification Based on Name

The algorithm makes use of name lists from three different sources: the US Census², the work of researchers Diane Lauderdale and Bert Kestenbaum³, and seven participating NAACCR registries. The first source includes only surnames, while the other two sources include both first names and surnames.

The Census list is based on the complete count of the 2000 census. This list contains all surnames occurring at least 50 times with percentages for white, black, American Indian/Alaska Native, Chinese, Japanese, Filipino, Korean, Asian Indian, Vietnamese, Other Asian, Hawaiian, Guamanian/Chamorro, Samoan, Other Pacific Islander, and Other, where at least 50% of the persons with these names are Asian or Pacific Islander. While dated 2002, to the best of our knowledge this was not released until 2008 at the earliest, and so was not incorporated into NAPIIA until version 1.2. As this list is population-based, it is considered the most definitive list available, and so is the first list against which names are compared. For a case coded 96, if more than 75% of the occurrences of the surname among all Asians occur within a specific Asian group, the case is recoded to that group. The same is true for cases coded 97 and Pacific Islanders.

Lauderdale and Kestenbaum published lists of surnames and first names strongly predictive of Chinese, Japanese, Korean, Filipino, Asian Indian, or Vietnamese race, based on an examination of 1.8 million Social Security applications for persons born in Asia before 1941. Collectively these are known as the “Lauderdale list”. The six race groups included on this list represent the largest Asian-American race groups and account for a large majority of the Asian-American population (91%, according to the 2000 census) and cancer incident case counts. Names were included on the list if at least 75% of the occurrences of the name were associated with a single one of the six groups and they occurred at least 4 times.

Finally, a list was developed derived from 80,000 cancer cases among Asians from 1997-2001 from seven NAACCR registries (Hawaii, Los Angeles, Louisiana, Illinois, Nevada, New York, Texas), applying the same criteria as for the Lauderdale list, and known as the “NAACCR list”. Surnames were deleted from the Lauderdale and NAACCR lists if less than 10% of the occurrences of the name in the 2000 Census were among Asian persons. Names were also deleted from the first name lists that obviously were not typically Asian^c.

A brief Hmong name list was supplied by Richard Yang of the Cancer Registry of Central California, who has extensive experience analyzing the Hmong population⁴; these names are evaluated at the same level as the Census list. Names for other Asian and Pacific Islander groups (e.g., Tongan), may be developed for future versions of NAPIIA.

Cases with a race code of 96 that were not indirectly identified or excluded in Step 3 based on birthplace are compared with the Census, Lauderdale and NAACCR lists in the following sequence^{d,e}. Upon attaining a match, the process is stopped and no further comparisons are made:

For males:

Table 4.1. Males	
M1. Check surname with Census surname list	
M2. Check surname with Lauderdale surname list	
M3. Check surname with NAACCR surname list	
M4. Check given name with Lauderdale given name list	
M5. Check given name with NAACCR given name list	

For females:^f

Table 4.2 Females	
F1. Check whether maiden name is blank:	
If blank:	If not blank:
F2a. Check surname with Census surname list	F2b. Check maiden name with Census surname list
F3a. Check surname with Lauderdale surname list	F3b. Check maiden name with Lauderdale surname list
F4a. Check surname with NAACCR surname list	F4b. Check maiden name with NAACCR surname list
F5a. Check given name with Lauderdale given name list	F5b. Check given name with Lauderdale given name list
F6a. Check given name with NAACCR given name list	F6b. Check given name with NAACCR given name list
	F7b. Check surname with Census surname list
	F8b. Check surname with Lauderdale surname lists
	F9b. Check surname with NAACCR surname list

Cases meeting none of these criteria will remain as a code 96 or 97.

Table 4.3 provides several examples on how the above rules are applied.

Table 4.3. Examples				
Sex	Name	Maiden Name	Assign to	Reason
F	Masako Smith	Nakamura	Japanese	Nakamura is on the Census surname list (rule F2b).
F	Shui Tong	Law	Chinese	Law is not on the Census surname list (<50% API) or Lauderdale surname list, but has a PPV of 0.89 for Chinese on the NAACCR list (rule F4b).
F	Maria Peralta	missing	Filipino	Peralta is not on the Census surname list (<50% API), but is on the Lauderdale surname list (rule F3a).
F	Gumti Chowdhury	missing	Asian Indian	Chowdhury is on the Census surname list, and while indicative of Asian (PPV=0.85), is not indicative of Asian Indian (PPV=0.62). It is not on the Lauderdale surname list, but has a PPV of 1.00 for Asian Indian on the NAACCR list (rule F4a).
F	Phuong Hang	Hua	Vietnamese	On the Census surname list, both Hang and Hua are strongly Asian, but not indicative of a specific group. Neither name is on the Lauderdale or NAACCR list. Phuong is on the Lauderdale list for given name (rule F5b).
M	Hyung Kim	n/a	Korean	Kim is on the Census surname list (rule M1).
M	Seong Moon	n/a	Korean	Moon is not on the Census surname list (<50% API) or the Lauderdale surname list but has a PPV of 0.88 for Korean on the NAACCR list (rule M3).
M	Byong Lee	n/a	Korean	Lee is not on the Census surname list (<50% API), Lauderdale surname list, or NAACCR surname list (ambiguous whether Korean or Chinese). Byong is on the Lauderdale given name list (rule M4).

The NAPIIA algorithm has been computerized and is available, with the name lists, on the NAACCR website. It runs as part of a SAS program that also calculates NHIA (NAACCR Hispanic Identification Algorithm). The two algorithms are independent and can be run singly, but are bundled together for convenience.

The SAS code produces detailed reports for each step of the process. These include listings of all records requiring manual review, listings of all records with their newly assigned NAPIIA code, frequency tables of newly assigned NAPIIA codes, and frequency tables of race vs. birthplace for birthplaces excluded

from the algorithm. Registries should review these reports to increase their understanding of nuances in local data that might suggest training issues, data quality and consolidation issues, potential for misclassification using indirect means, or other local effects.

Quality Evaluation

The New York, Louisiana and Los Angeles registries evaluated the quality of version 1.1 by setting all cases with known Asian race to 96, and seeing if the algorithm returned the original race (The Hawaii registry later also performed this evaluation, but its results are not included in the summary tables below). The largest number of cases were assigned based on birthplace, and these were also the most accurate. The second-largest number of cases was assigned using the Lauderdale surname list, and these were the second-most accurate. Generally, as the algorithm proceeds, both the number of cases assigned and the accuracy decreases.

This exercise will be repeated for the next major algorithm update (version 2.0). For version 1.2, the accuracy of the Census surname list is attested to in a study of 1.9 million enrollees of a national health plan conducted by Elliott et al.⁵

Table Q1. Quality Evaluation Results for Males		
Description	N	% correct
Birth place	10,395	98%
Match surname against Lauderdale list	3,788	93%
Match surname against NAACCR list	257	87%
Match first name against Lauderdale list	338	83%
Match first name against NAACCR list	129	81%
TOTAL	14,907	96%

Table Q2. Quality Evaluation Results for Females without Maiden Name		
Description	N	% correct
Birth place	10,081	99%
Match surname against Lauderdale list	3083	92%
Match surname against NAACCR list	221	88%
Match first name against Lauderdale list	414	87%
Match first name against NAACCR list	169	83%
TOTAL	13,968	97%

Table Q3. Quality Evaluation Results for Females with Maiden Name		
Description	N	% correct
Birth place	1,790	97%
Match maiden name against Lauderdale list	416	88%
Match maiden name against NAACCR list	40	85%
Match first name against Lauderdale list	54	93%
Match first name against NAACCR list ^g	25	68%
Match surname against Lauderdale list	67	81%
Match surname against NAACCR list	18	83%
TOTAL	2,410	95%

Note that these results are not truly representative of cases actually coded as 96. They represent cases where the race was already known, and are more likely to have a known birthplace, and less likely to have a highly unusual name, than cases actually coded as 96. Thus, the percent correct is probably higher than will be seen in practice. Still, these results establish an overall confidence in NAPIIA.

References

1. Johnson CH (ed.), *SEER Program Coding and Staging Manual 2004, Revision 1*. National Cancer Institute, NIH Publication number 04-5581, Bethesda, MD.
2. Falkenstein MR. The Asian and Pacific Islander Surname List as Developed from Census 2000. In *Joint Statistical Meetings*, New York, 2002.
3. Lauderdale DS and Kestenbaum B. Asian American Ethnic Identification by Surname. *Population Research and Policy Review* 2000;19: 283-300.
4. Mills PK, Yang RC, Riordan D. Cancer Incidence in the Hmong in California, 1988-2000. *Cancer* 2005; 104: 2969-2974.
5. Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N. Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities. *Health Serv Outcomes Res Method* 9: 69-83; 2009.

Notes on Algorithm Development

- a. Birthplace of Laos was originally coded to Laotian, but upon greater awareness of the Hmong population in the US, many of whom were born in Laos, this was removed.
- b. Table 3.2 was the most extensively discussed and debated element of the algorithm. While it results in fewer recodes than if it were not applied, the number of US residents born in these places is small, and so the overall effect on the power of the algorithm is modest.
- c. A typical example is the surname Kennedy, which the Lauderdale list defines as Japanese. A single missionary or diplomatic family with this surname who had four children while based in Japan could account for this finding, but this has little bearing on whether a cancer patient with this surname with a race code of 96 should also be considered Japanese. (If the Lauderdale list had been based on a sample size considerably larger than 4, this problem could have been minimized). Many surnames were deleted using this criterion (roughly 20% of the total), but only 2% of the cases with code 96 were affected. Lacking a master tally of first names by race, we had to rely on common sense when pruning the list. William and Claire are typical examples of names that were removed.
- d. All three lists add value to the algorithm. The Census list excludes names like Park and Moon that are more common among non-Asians, but that are predictive conditional on being Asian, as with code 96. In addition, the Lauderdale and NAACCR lists reflect an older cohort and a cancer cohort, respectively. These lists include names that are not predictive of a specific Asian group population-wide, but that are predictive for an older population. An example is the surname Lo. In the 2000 Census, this name was indeterminate as to whether Chinese or Vietnamese. In both the Lauderdale and NAACCR lists, however, this name is highly predictive of Chinese, reflecting the fact that the Vietnamese population in the US is younger than the Chinese population.
- e. The original algorithm included a reverse name check. This step enabled a check of a first name with the surname field and vice versa under the assumption that these names could easily get reversed in a medical record, particularly where some Asian cultures present themselves using their surname first. A lack of familiarity with Asian names would minimize the chance that these would get corrected. This assumption was checked by testing in New York and Louisiana. All persons with a known Asian race in the registry were recoded to an Asian NOS race to determine whether NAPIIA correctly re-assigned them to the same specific Asian race. Overall, NAPIIA worked very well, except for the reverse name checks, where the misclassification rate was very high. As the reverse name check caused more problems than it resolved, the decision was made to eliminate a reverse name check.

f. Originally for Step 4, the order of precedence for women was maiden name, then first name, then surname. However, it was discovered through the empirical testing described above, that when the maiden name field was blank, matching the surname was more accurate than the given name. The algorithm was revised accordingly.

g. The result of the NAACCR first name list where a maiden name is present (17 out of 25 correct) is lower than the target threshold of 75% and less accurate than the steps that follow. However, the sample size is very small and this step should be considered in combination with the result of the NAACCR first name list where maiden name is absent (83% correct).