

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

*October 19, 2007*



**Editors:**

**NAACCR Asian/Pacific Islander Work Group**

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

**Suggested Citation:**

NAACCR Asian Pacific Islander Work Group. *NAACCR Asian Pacific Islander Identification Algorithm [NAPIIA v1]*. Springfield (IL): North American Association of Central Cancer Registries. October 2007.

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.

## **Work Group Members**

Francis P. Boscoe, Chair

Peg Balcius

Cheryll Cardinez

Vivien W. Chen

Catherine Grafel-Anderson

Scarlett Gomez

Michael Green

Holly L. Howe, Chair 2004-05

Mei-chin Hsieh

Betsy Kohler

Sandy Kwong

Andy Lake

Lihua Liu

Barry Miller

Steve Schwartz

Maria J. Schymura

Melanie Williams

## **NAACCR Asian Pacific Islander Identification Algorithm (NAPIIA) v1**

### *Summary*

The NAACCR Asian Pacific Islander Identification Algorithm version 1 (NAPIIA v1) uses a combination of NAACCR variables to classify cases directly or indirectly as Asian Pacific Islander for analytic purposes. This version of the algorithm is focused on coding cases with a race code of Asian NOS (race code 96) to a more specific Asian 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 race groups (Chinese, Japanese, Vietnamese, Korean, Asian Indian, Filipino, Thai, and Cambodian). 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). Future versions of NAPIIA will incorporate Pacific Islanders and will potentially incorporate name lists for Thai, Cambodian, and Laotians.

The algorithm uses the following NAACCR standard variables:

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

*Detailed NAPIIA v1 Guidelines*

**Step 1. Identify cases containing race code 96**

**1.1. Single race code of 96.** All cases with a Race 1 code (data item 160) of 96 and no additional race codes will be identified and retained for Steps 3 and 4 of the algorithm (Table 1.1). For these cases, the codes for Race 2 through Race 5 (data items 161-164) must be blank or 88.

<b>Race 1 Code</b>	<b>Category</b>
96	Other Asian, Asian NOS, Oriental NOS

**NOTE:** Race code 99 (unknown) is not an acceptable code when used in combination with any other race code. If one race code is 99, then all race codes must be 99.

**1.2. Single race code of 97.** This section reserved for future editions of NAPIIA<sup>a</sup>.

**1.3 Race code of 96 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. 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. 46-50<sup>1</sup>.

<b>Table 1.3 Multiple race code scenarios involving code 96</b>	
<b>Scenario</b>	<b>Action</b>
1.3.1. One race code is 04-32, one race code is 96; others are blank or 88.	The 04-32 takes precedence. Treat as a single race case. Go to step 2.
1.3.2. More than one race code is 04-32; one race code is 96; others are blank or 88.	The contents of Race 1 are output, and the case is flagged for manual review.
1.3.3. One race code is 01; one race code is 96; others are blank or 88.	The 96 code takes precedence. Go to step 3.
1.3.4. One or more race codes is 02-03; one race code is 96; others are blank or 88.	The contents of Race 1 are output, and the case is flagged for manual review.
1.3.5. 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.
1.3.6. Any multiple race combination involving code 96 not listed above	The contents of Race 1 are output, and the case is flagged for manual review.

**NOTE:** As of 2006, very few cancer cases are reported to central registries with more than one race, with the number of cases in the entire United States well below 0.5% of total cases reported. If reporting of multi-race cases becomes more common in the future, then this step should be re-evaluated for continued appropriateness and validity.

## Step 2. Directly code cases not containing code 96.

### 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, 97, 98, and 99 in Race 1 (data item 160) (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.

<b>Code</b>	<b>Category</b>
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
09	Asian Indian, Pakistani
10	Vietnamese
11	Laotian
12	Hmong
13	Kampuchean
14	Thai
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
97	Pacific Islander, NOS
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.

**NOTE:** As of 2006, very few cancer cases are reported to central registries with more than one race, with the number of cases in the entire United States well below 0.5% of total cases reported. If reporting of multi-race cases becomes more common in the future, then this step should be re-evaluated for continued appropriateness and validity.

<b>Scenario</b>	<b>Action</b>
2.2.1. One race code is 01, one race code is 02-32 or 97; others are blank or 88.	The 02-32 or 97 takes precedence.
2.2.2. 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 is applied only to single race persons that have a code of 96 on NAACCR Standard data item 160 as identified in Steps 1.1, or certain multiple race persons identified in Step 1.3.

If a person has a birthplace (data item 250) 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<sup>b</sup>. These persons have a high probability of being of the specific Asian or Pacific Islander race groups identifiable in NAPIIA v1.

<b>Code</b>	<b>Birthplace</b>	<b>Race</b>
681, 682, 683, 684, 686	China, Taiwan, Hong Kong, Macao	Chinese
133, 134, 693	Nampo-Shoto, Ryukyu Islands, Japan	Japanese
675	Philippines	Filipino
695	Korea, North Korean, South Korean	Korean
639, 641	India, Pakistan	Asian Indian, Pakistani
665	Vietnam	Vietnamese
651	Thailand	Thai
663	Cambodia, Kampuchea	Cambodian or Kampuchean

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).

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

It has been observed that a name could 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 recorded in NAPIIA version 1. This issue will be further considered in subsequent versions of NAPIIA<sup>c</sup>.

#### **Step 4. Indirect Identification Based on Name**

Lauderdale and Kestenbaum have 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<sup>2</sup>. 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 countries (PPV  $\geq$  .75) and they occurred at least 4 times. This list was supplemented by a list 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, known as the “NAACCR list”<sup>d</sup>. A brief Hmong name list was supplied by Richard Yang of the Cancer Registry of Central California, who has extensive experience analyzing the Hmong population<sup>3</sup>; these names are also considered part of the NAACCR list. Names for other Asian groups (e.g., Thai), as well as Pacific Islanders, are expected to be added in 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 Lauderdale and NAACCR lists in the following sequence<sup>e</sup>. Upon attaining a match, the process is stopped and no further comparisons are made:

For males:

<b>Table 4.1. Males</b>	
M1. Check <b>surname</b> with Lauderdale <b>surname</b> lists	
M2. Check <b>surname</b> with NAACCR <b>surname</b> List (PPV > =.75)	
M3. Check <b>given</b> name with Lauderdale <b>given</b> name lists	
M4. Check <b>given</b> name with NAACCR <b>given</b> Name List (PPV > =.75)	

For females<sup>f</sup>:

<b>Table 4.2 Females</b>	
F1. Check <b>maiden</b> name with Lauderdale <b>surname</b> lists	
F2. Check <b>maiden</b> name with NAACCR <b>surname</b> List (PPV > =.75)	
F3. Check whether <b>maiden</b> name is blank.	
F3a. If <b>maiden</b> name is blank:	F3b. If <b>maiden</b> name is not blank:
F4a. Check <b>surname</b> with Lauderdale <b>surname</b> lists	F4b. Check <b>given</b> name with Lauderdale <b>given</b> name lists
F5a. Check <b>surname</b> with NAACCR <b>surname</b> List (PPV > =.75)	F5b. Check <b>given</b> name with NAACCR <b>given</b> Name List (PPV > =.75)
F6a. Check <b>given</b> name with Lauderdale <b>given</b> name lists	F6b. Check <b>surname</b> with Lauderdale <b>surname</b> lists
F7a. Check <b>given</b> name with NAACCR <b>given</b> Name List (PPV > =.75)	F7b. Check <b>surname</b> with NAACCR <b>surname</b> List (PPV > =.75)

Cases meeting none of these criteria will remain as a code 96, Asian NOS.

Table 4.3 provides several examples on how the rules are applied:

<b>Table 4.3. Examples</b>				
<b>Sex</b>	<b>Name</b>	<b>Maiden Name</b>	<b>Assign to</b>	<b>Reason</b>
F	Masako Smith	Nakamura	Japanese	Nakamura is on the Lauderdale surname list (rule F1).
F	Shui Tong	Law	Chinese	Law is not on the Lauderdale surname list, but has a PPV of 0.89 for Chinese on the NAACCR list (rule F2).
F	Maria Peralta	missing	Filipino	Peralta is on the Lauderdale surname list (rule F4a).
F	Gumti Chowdhury	missing	Asian Indian	Chowdhury is not on the Lauderdale surname list, but has a PPV of 1.00 for Asian Indian on the NAACCR list (rule F5a).
F	Phuong Hang	Hua	Vietnamese	Neither Hang nor Hua is on the Lauderdale surname list, and they both have low PPVs on the NAACCR list (ambiguous whether Vietnamese or Chinese), but Phuong is on the Lauderdale list for given name (rule F4b).
M	Hyung Kim	n/a	Korean	Kim is on the Lauderdale surname list (rule M1).
M	Seong Moon	n/a	Korean	Moon is not on the Lauderdale surname list but has a PPV of 0.88 for Korean on the NAACCR list (rule M2).
M	Byong Lee	n/a	Korean	Lee is not on the Lauderdale surname list, has a low PPV on the NAACCR list (ambiguous whether Korean or Chinese), but Byong is on the Lauderdale given name list (rule M3).

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 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 the algorithm 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 were 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.

Table Q1. Quality Evaluation Results for Males			
Step	Description	N	% correct
	Birth place	10,395	98%
F4a	Match surname against Lauderdale list	3,788	93%
F5a	Match surname against NAACCR list	257	87%
F6a	Match first name against Lauderdale list	338	83%
F7a	Match first name against NAACCR list	129	81%
	TOTAL	14,907	96%

Table Q2. Quality Evaluation Results for Females without Maiden Name			
Step	Description	N	% correct
	Birth place	10,081	99%
F1	Match surname against Lauderdale list	3083	92%
F2	Match surname against NAACCR list	221	88%
F3	Match first name against Lauderdale list	414	87%
F4	Match first name against NAACCR list	169	83%
	TOTAL	13,968	97%

Table Q2. Quality Evaluation Results for Females with Maiden Name			
Step	Description	N	% correct
	Birth place	1,790	97%
F1	Match maiden name against Lauderdale list	416	88%
F2	Match maiden name against NAACCR list	40	85%
F4b	Match first name against Lauderdale list	54	93%
F5b	Match first name against NAACCR list <sup>g</sup>	25	68%
F6b	Match surname against Lauderdale list	67	81%
F7b	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 v1.

## ***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. Lauderdale DS and Kestenbaum B. Asian American Ethnic Identification by Surname. *Population Research and Policy Review* 2000;19: 283-300.
3. Mills PK, Yang RC, Riordan D. Cancer Incidence in the Hmong in California, 1988-2000. *Cancer* 2005; 104: 2969-2974.

## ***Notes on Algorithm Development***

a. Pacific Islanders were added to the algorithm between July 2006-February 2007, but given the paucity of the available name lists, it was decided to remove the PI portion until better name lists could be developed (specifically, we are awaiting the results of a research project led by Myles Cockburn).

b. 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.

c. 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. An exception is the Hawaii registry, where Hawaii is the principal birthplace, meaning that the algorithm lacks power here. However, given the unique and complex race characteristics of Hawaii, Hawaii registry staff prefer to manually review their small number of race code 96 and 97 cases rather than depend on the algorithm results.

d. Anglican given names on the NAACCR list that exceed the case count and PPV thresholds (e.g. William, Claire) were found to lead to inaccurate reassignment. We thus evaluated the frequency of names on the whole database and the proportion that occurs in Asians with respect to non-Asians and eliminated those that were predominantly non-Asian.

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 results of the NAACCR first name list where a maiden name is present (17 out of 25 correct) are a bit lower than our target threshold of 75% and are less accurate than the steps that follow. However, the sample size is very small here and step F4b should be considered in combination with step F4, where the correct classification rate was 83%.