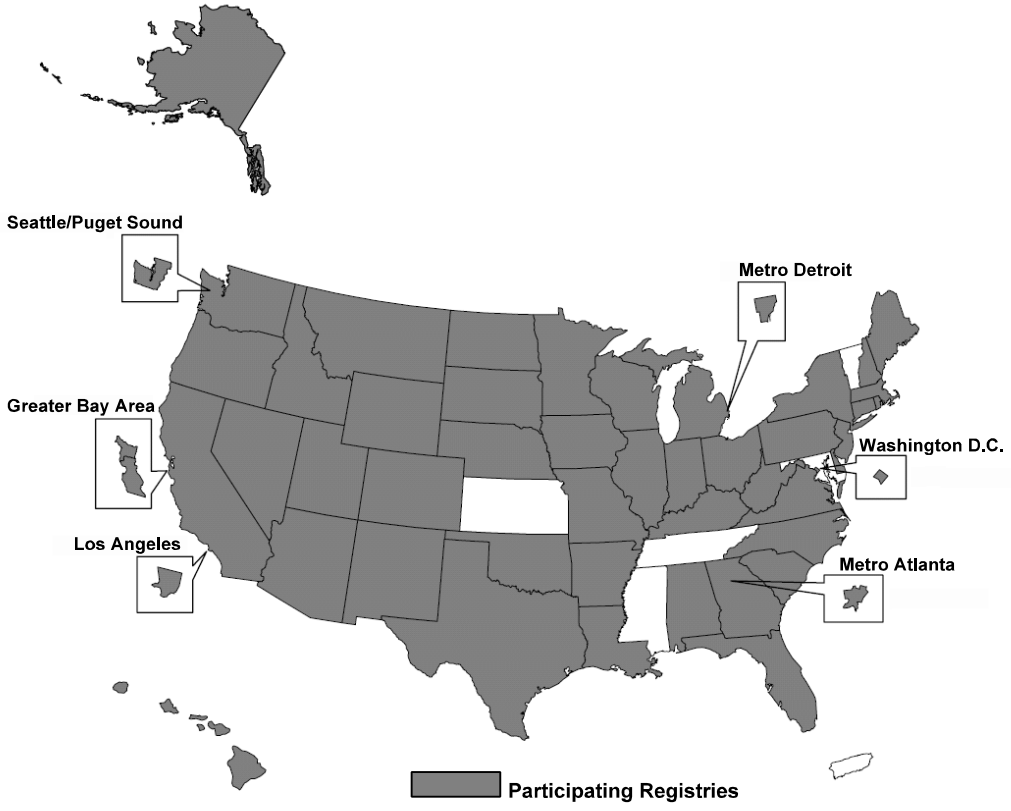
Appendix F

Description of the NAACCR Hispanic Identification Algorithm (NHIA)



Appendix F: NAACCR Best Practice Guideline for Hispanic Identification

Extracted from the NAACCR Report:

NAACCR Expert Panel in Hispanic Identification. *Report of the NAACCR Expert Panel on Hispanic Identification 2003*. Springfield (IL): North American Association of Central Cancer Registries, October 2003.

Direct Identification of Hispanic Persons

Ideally, the best approach to identify cancer cases who are Hispanic is a direct one. Registries need to promote among reporting facilities the importance of documenting all race and Hispanic ethnicity identifiers in the medical record. The existing registry process for abstracting race and Hispanic identifiers, including birth place information and maiden name, needs to be reviewed, assessed, and improved, capturing all available information from the medical record and abstracting it to the cancer reporting form. This process should be incorporated into all training and education programs. Registries must be cautious about relying on facilities to assign a code related to Hispanic ethnicity that employs all the same criteria as the central registry. For example, unless the central registry is assured that a facility is using and following the central registry's standard surname algorithm program or list, it should not assume that a code of 7 on data element 190 is a valid code. Similarly, an assignment of a code of 0 to data element 190 may also not be assigned in a reliable or valid manner, unless the facility is carefully following the protocols of the central registry.

For cases diagnosed in 2000 and later, the registry must establish rules for handling inconsistent race and Hispanic ethnicity identification. The questions must be answered as to whether these are true multi-race cases or true errors/ inconsistencies. While a person can be multi-race, they cannot be both Hispanic and non-Hispanic.

Indirect Identification of Hispanic Persons

Sometimes, despite best efforts to obtain complete information directly from the medical record, information is not available and is reported to the cancer registry as a missing data item. With regard to Hispanic ethnicity, some cancer registries have found it necessary to rely on indirect methods to populate this data element. The registries often have significant numbers or proportions of Hispanic populations in their jurisdiction. They have needed to develop alternate approaches to enhance Hispanic identification which include reliance on death certificates, surname and maiden name matching algorithms, birth place, special studies, physician follow-up, and linkage with other data sources that enhance Hispanic identification.

Based on the NAACCR survey of all registries and the empirical evaluation of representatives from states that produce cancer incidence data for the Hispanic population in their registry area, the best practice guideline is that all registries follow the NAACCR Hispanic Identification Algorithm (NHIA). This can be accomplished in one of three ways (or combination): 1) following the step-by-step guidelines enumerated below; 2) following the diagram described in Figures 3-5 of the original report (not reproduced here), or 3) applying a computerized algorithm of these guidelines (a SAS version is available for download from the NAACCR web site). You also will need the 1990 Hispanic Surname list from the U.S. Census (also available on the NAACCR website). The NHIA guideline was adapted from the Illinois State Cancer Registry Hispanic Algorithm. The Colorado Central Cancer Registry developed the SAS version of the computerized NHIA algorithm.

The NAACCR Hispanic Identification Algorithm (NHIA)

The NAACCR Hispanic Identification Algorithm (NHIA) uses a combination of NAACCR variables to directly or indirectly classify cases as Hispanic for analytic purposes. It is possible to separate Hispanic ancestral subgroups (e.g., Mexican) when indirect assignment results from birth place information but not from the surname match. The algorithm uses the following NAACCR standard variables: Spanish/Hispanic Origin (item 190), Name-Last (item 2230), Name-Maiden (item 2390), Birthplace (item 250), Race 1 (item 160), and Sex (item 220).

Summary

A person is classified as **Hispanic** or **non-Hispanic** using NHIA through either direct or indirect identification.

Direct Identification. Cases reported as Spanish/Hispanic Origin(item 190):

1-Mexican;
2-Puerto Rican;
3-Cuban;
4-South or Central American (except Brazil);
5-Other specified Spanish/Hispanic origin (includes European);
6- Spanish, NOS, Hispanic, NOS, Latino, NOS.

Indirect Identification. Cases reported as Spanish/Hispanic Origin:

0-non-Spanish/non-Hispanic;
7- Spanish surname only;
9-Unknown whether Spanish.

Persons are excluded in the indirect identification process if they are of Filipino, Native American, or Hawaiian race or when they are born in certain countries (see below for specific list) (thus classified as **non-Hispanic**).

Persons are included in the indirect identification process if they are born in birthplaces with a high probability of Hispanic ethnicity (see below for specific list).

Persons are also included as **Hispanic** when they are male cases with **heavily Hispanic** last names; female cases with **heavily Hispanic** maiden names; female cases with missing maiden names and **heavily Hispanic** last names; female cases with **generally Hispanic, moderately Hispanic or occasionally Hispanic** maiden names and **heavily Hispanic** last names.

After applying NHIA, cases not classified as Hispanic are classified as **non-Hispanic**.

Detailed NHIA Guidelines

Step 1. Evaluate NAACCR Data Element 190 Codes.

Step 1.1 Spanish/Hispanic Origin Data Element (NAACCR Data Element 190)	
Code	**Category**
1	Mexican (includes Chicano)
2	Puerto Rican
3	Cuban
4	South or Central American (except Brazil)
5	Other specified Spanish/Hispanic origin (includes European)
6	Spanish, NOS; Hispanic, NOS; Latino, NOS (NOS- Not otherwise specified)

For NAACCR standard data element 190, all cases reported by reporting facilities as Spanish/Hispanic origin, with codes 1, 2, 3, 4, 5 and 6 the codes for **Hispanic**. These codes are detailed in the table for Step 1.1. This step represents the direct identification component of the NAACCR Hispanic Identification Algorithm (NHIA).

Step1.2 Spanish/Hispanic Origin Data Element (NAACCR Data Element 190)	
Code	**Category**
0	Non-Hispanic
7	Surname only
9	Unknown

The indirect identification component involves cancer cases reported as Spanish/Hispanic origin data element codes 0, 7 and 9 (see table of step 1.2) for the NAACCR standard data element 190. The goal is to classify these cases as Hispanic or non-Hispanic based on an evaluation of the strength of the birthplace, race, and/or surname associations with Hispanic ethnicity status.

Step 2. Assign Last and Maiden Surnames According to the 1990 Census Bureau Spanish Surname List

Match last and maiden surnames on the cancer registry database to the 1990 U.S. Bureau of the Census Spanish surname list assigning codes to registry surnames matching those from the census bureau study sample. At this point, all surnames on the cancer registry database will **not** have been coded using the census bureau surname list. Surnames not appearing in the census bureau study sample (indeterminate) or missing surnames will not have been coded during the match. Assign code 6000 to surnames not on the census bureau study sample list and 9000 to missing surnames. All last and maiden surnames on the registry database should now have been assigned one of the codes shown in the table for Step 2.

Step 2. Surname Codes						
Heavily Hispanic	101	102	105	110	115	125
Generally Hispanic	201	202	205	210	215	225
Moderately Hispanic	301	302	305	310	315	325
Occasionally Hispanic	-	-	405	410	415	425
Rarely Hispanic	5001	5005	5010	5025	5100	5500
Indeterminate	-	-	-	-	-	6000
Missing	-	-	-	-	-	9000

Step 3. Filter Cases for Indirect Identification Based on Birthplace

3.1 Some cases are assigned to Hispanic ethnicity based on birthplace. Cases born in birthplaces associated with a high prevalence of Spanish surnames but a low probability of Hispanic ethnic status (see table of Step 3.1). Anyone with a birthplace listed in the following table would be coded to 0, non-Hispanic.

Step 3.1. Birthplaces Associated with Prevalence of Spanish Surnames but Low Probability of Hispanic Ethnicity	
Code	**Birthplace**
100, 102, 109	Atlantic/Caribbean area **excluding** Cuba , Dominican Republic, and Puerto Rico
110	Panama Canal
120-137	Pacific Area
341	Brazil
331	Guyana
332	Suriname
333	French Guyana
400-441; 445-499	Europe including Portugal
675	Philippines

3.2 In general, those cases born in birthplaces shown in the table for Step 3.2 have high probabilities of being Hispanic. Although reporting guidelines encourage review of birthplace information when reporting Spanish/Hispanic origin, this step seeks to identify those cases missed during the reporting process. Remaining cases born in birthplaces with high probability of Hispanic ethnicity are classified **Hispanic** using NHIA.

Step 3.2. Birthplaces with High Probability of Hispanic Ethnicity					
Code	**Birthplace**	**NHIA**	**Code**	**Birthplace**	**NHIA**
101	Puerto Rico	2	265	Latin America NOS	4
230	Mexico	1	300	South America	4
241	Cuba	3	311	Colombia	4
243	Dominican Republic	4	321	Venezuela	4
250	Central America	4	345	Ecuador	4
251	Guatemala	4	351	Peru	4
252	Belize	4	355	Bolivia	4
253	Honduras	4	361	Chile	4
254	El Salvador	4	365	Argentina	4
255	Nicaragua	4	371	Paraguay	4
256	Costa Rica	4	375	Uruguay	4
257	Panama	4	443	Spain (including Canary Islands, Balearic Island, and Andorra).	5

Step 4. Exclude Cases from Indirect Identification Based on Race

Cases reported as race codes 03-Native American, 06-Filipino or 07-Hawaiian are eliminated from indirect identification as these race groups often have Spanish surnames but are generally not of Hispanic ethnicity.

Step 5. Exclude Cases from Indirect Identification Based on Surname Codes (by Sex)

For male cases, those with last names coded as rarely Hispanic, indeterminate, or unknown (5001, 5005, 5010, 5025, 5100, 5500, 6000 or 9000) are eliminated from indirect identification. For female cases,

those with maiden names coded as 5001, 5005, 5010, 5025, 5100, 5500 or 6000 are eliminated from indirect identification. (If a large percentage of maiden names are missing, last name is used for indirect identification purposes.) For remaining female cases, those with maiden names codes as 9000 and last names coded as 5001, 5005, 5010, 5025, 5100, 5500, 6000 or 9000 are eliminated from indirect identification. The cases meeting this criterion would be coded to 0, non-Hispanic.

Step 6. Identify Cases with Heavily Hispanic Surnames (by Sex)

For the remaining cases, males with last names coded to heavily Hispanic (125, 115, 110, 105, 102 or 101) are classified **Hispanic**. Those females with available maiden names coded to heavily Hispanic (125, 115, 110, 105, 102 or 101) are classified **Hispanic**. Females with missing maiden names (9000) or maiden names classified as generally, moderately or occasionally Hispanic (201, 202, 205, 210, 215, 225, 301, 302, 305, 310, 315, 325, 405, 410, 415, 425) and whose last names are heavily Hispanic (coded 125, 115, 110, 105, 102 or 101) are classified **Hispanic**. The remaining cases are classified as **non-Hispanic.**

Step 7. Save the results of NHIA as a separate data element.

Step 7. NHIA Data Element	
Code	**Category**
0	Non-Hispanic
1	Mexican, by birthplace or other specific identifier
2	Puerto Rican, by birthplace or other specific identifier
3	Cuban, by birthplace or other specific identifier
4	South or Central American (except Brazil), by birthplace or other specific identifier
5	Other specified Spanish/Hispanic origin (includes European), by birthplace or other specific identifier
6	Spanish, NOS; Hispanic, NOS; Latino, NOS (NOS-Not otherwise specified)
7	NHIA surname match only

The results of NHIA need to be recorded or saved as a separate data element. The same coding values as for NAACCR standard data element 190 should be used, as shown in the table for step 7. The one exception is that no missing codes will be allowed, because at the conclusion of step 6 of NHIA, if a case has not been identified as Hispanic, it will be coded to 0, non-Hispanic. The new variable has been proposed to the NAACCR Uniform Data Standards committee to become a standard variable in 2006, at the next major revision of NAACCR standards. Until that time it will be transmitted to NAACCR in column number 1028 in version 9.1 of the NAACCR Data Exchange Layout.

Procedural Considerations

1. For data element 190, Spanish/Hispanic Origin, neither a reporting source nor a computer system should default to a non-Hispanic identification. If any default is used, it should be to the Hispanic ethnicity unknown (code 9 on NAACCR data standard element 190).

2. Central registries should ignore all Hispanic case reports that have been coded by a reporting facility with the value of "7", surname only, for data element 190 UNLESS the central registry is assured that the facility is using the same surname matching algorithm as the central registry. If the hospital is not, treat all these cases as a "9", unknown if Spanish/Hispanic.

3. Rate calculations should ensure that the numerator matches the denominator for both race and Hispanic ethnicity. The consensus of the group is to report Hispanic rates for all race groups combined (Hispanic, All Races). This is with the understanding that for persons of unknown Hispanic ethnicity who also have a race or birthplace as the Philippines or a race that is Hawaiian, a surname matching algorithm will not be used to identify them as Hispanic.

4. Run the algorithm for all cases with data element #190, Spanish/Hispanic Origin coded to either a 0 (non-Spanish; non-Hispanic), a 7 (report source states surname only basis) or a 9 (unknown whether Spanish). If a registry has objective criteria or reasons to demonstrate that inclusion of persons coded as 0 (non-Spanish; non-Hispanic) causes an over-identification of Hispanic persons, then it may be acceptable to run the algorithm only on cases coded to either a 7 or a 9. However, this decision must be based on valid, scientific assessments with written documentation of results. This information will need to be supplied to NAACCR with a file submitted in response to a Call for Data.

5. Make sure that the results of the entire Hispanic identification process are stored in the registry database and updated with new information. As an alternative, Hispanic ethnicity can be automatically derived each time a data use file is created using relevant data elements.