

**NAACCR Guideline for
Enhancing Hispanic-Latino
Identification:
Revised NAACCR
Hispanic/Latino Identification
Algorithm [NHIA v2.2.1]**

Revised September 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 Guideline for Enhancing Hispanic/Latino Identification: Revised NAACCR Hispanic/Latino Identification Algorithm [NHIA v2.2.1]*. Springfield (IL): North American Association of Central Cancer Registries. September 2011.

Cooperative Agreement Number U75/CCU523346 from CDC provided funds 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 (chair)
New York State Cancer Registry
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 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

Versions 1.0 and 2.0 of this document were originally developed by the NAACCR Latino Research Group, chaired by Holly Howe, former Executive Director of NAACCR.

Table of Contents

Background	5
What's New in Version 2.2 and 2.2.1	6
NAACCR Guideline for Hispanic/Latino Identification	6
Direct Identification of Hispanic/Latino Persons	6
Indirect Identification of Hispanic/Latino Persons	7
The NAACCR Hispanic/Latino Identification Algorithm (NHIA) version 2.2.1)	7
Summary	8
Detailed NHIA v2.2.1 Logic	9
Step 1. Evaluate NAACCR Data Element 190 Codes.	9
Step 2. Filter Cases for Indirect Identification Based on Birthplace	9
Step 3. Exclude Cases from Indirect Identification Based on Race	10
Step 4. Filter Cases for Indirect Identification Based on County of Residence	11
Step 5. Indirect Identification Based on Surname Codes (by Sex)	11
Step 6. Save the results of NHIA v2 as a separate data element.	11
Procedural Considerations	12
References	13
Appendix A. Sensitivity and Specificity of Heavily Hispanic Surnames based on the 1990 Census Spanish Origin Research file	14
Appendix B. Inflation of Hispanic cases using 1990 Census heavily Hispanic surnames to identify Hispanics as a function of the proportion of Hispanics in the population	15

Background

NAACCR convened an Expert Panel in 2001 to develop a best practices approach to Hispanic/Latino identification. Representatives were selected from registries that serve regions of the largest numbers of Hispanic/Latino populations in the United States. [1] The purpose of this activity was to evaluate the various methods and to determine whether a recommendation for one approach/ method was feasible among the various central cancer registries, considering the various Hispanic/Latino populations in the different geographic areas. A number of issues had to be considered in developing a best practice for Hispanic/Latino identification:

- No gold standard exists for comparison of cancer incidence rates for Hispanic/Latino populations.
- Cancer risks vary among subgroups of Hispanic/Latino population by country of origin.
- Persons of specific Hispanic/Latino origins (e.g., Mexican, Cuban, Puerto Rican, etc.) do not randomly occur across U.S. geographies. They cluster by geographic area.
- The age structure of the Hispanic/Latino population could vary for various Hispanic/Latino populations, as well as their length of residence and acculturation in the United States (related to risk).
- Race identification by Hispanic/Latino ethnicity varies and may vary regionally (i.e., white, black, or other race).
- Hispanic surname algorithms may not distinguish between Hispanic/Latino persons and persons of Portuguese, Italian, or Filipino descent.
- The responses to Hispanic/Latino origin questions have been inconsistent in self-reported information reported in the scientific literature.
- Information released from the 2000 U.S. Census suggests that annual population estimates used since the 1990 Census were not accurate and the differences were sufficiently large to affect the computation of rates for Hispanic/Latino populations.

The resulting approach to enhance Hispanic/Latino identification, the NAACCR Hispanic Identification Algorithm (NHIA), was computerized and released for use by central cancer registries in 2003. Further, the panel determined that NHIA was appropriate for application to cases diagnosed from 1995 forward. Application of the method for the years prior to 1995 was feasible, but the panel suggested that each registry should determine its appropriateness for these earlier years.

Cancer Incidence in U.S. Hispanics/Latinos, 1995-2000 (CIUSHL) was published in December 2003. [2] This NAACCR monograph included cancer incidence information for more than 85% of the total U.S. Hispanic/Latino population; but only 45% of the non-Hispanic white and 47% of the non-Hispanic black populations. In the NAACCR 2004 call for data, all registries were asked to run NHIA and submit their results for evaluation. [3] NHIA results were also submitted in the 2005 Call for Data and Hispanic/Latino rates were published as Volume IV of the monograph, *Cancer in North America, 1998-2002 [CINA]* [4]. Since then, Hispanic/Latino rates have been fully incorporated into *CINA*.

The development and testing of NHIA, and the resulting first volume of *CIUSHL*, was based on the experience of the states with the largest populations of U.S. Latino populations. When it was applied to states with smaller Latino populations, several issues emerged related to

over-identification and the positive predictive values of indirect identification using surname alone in areas with a low frequency of Latino populations. [4] Based on some state-specific analyses, several registries made suggestions to improve the accuracy of the surname-matching portion of the algorithm. These suggestions were reviewed and evaluated by the Latino Research Work Group, a group that evolved from the original Expert Panel on Hispanic identification. The resulting modifications were incorporated into the NAACCR Hispanic/Latino Identification Algorithm version 2 [NHIA v2], released in 2005. A small number of minor changes were made to the algorithm in 2008 by the NAACCR Asian/Pacific Islander Work Group, and released as version 2.1. Additional minor changes were made in 2009 by the same group, restyled the Race and Ethnicity Workgroup, as described in the following section.

What's New in Version 2.2 and 2.2.1

1. Belize has been removed from the list of countries likely to be Hispanic

As a minority of persons born in Belize identify as Hispanic, this is not considered predictive.

2. A comment in the program code has been added warning that the same case can be coded as both Hispanic and non-Hispanic under rare circumstances

This can occur among persons with multiple tumors diagnosed while residing in different counties, where one county is less than 5% Hispanic, one county is greater than 5% Hispanic, and Option 1 or Option 2 is selected.

3. Version 2.2.1: The algorithm is now compatible with both NAACCR Record Layout Versions 11.3 and 12.

4. In September, 2011, minor clarifications were made to this documentation without any changes to the algorithm itself.

NAACCR Guideline for Hispanic/Latino Identification

Direct Identification of Hispanic/Latino Persons

Ideally, the best approach to identify cancer cases who are Hispanic/Latino is a direct one. Registries need to promote among reporting facilities the importance of documenting all race and Hispanic/Latino ethnicity identifiers in the medical record. The existing registry process for abstracting race and Hispanic/Latino 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/Latino 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 not have been performed in a reliable or valid manner, unless the facility is carefully following the protocols or procedures established by the central registry.

For cases diagnosed in 2000 and later (when multiple race codes for each case were allowed), the registry must establish rules for handling inconsistent race and Hispanic/Latino ethnicity identification. The questions must be answered as to whether these are true multi-race cases or errors/ inconsistencies. While a person can be multi-race, he/she cannot be both Hispanic and non-Hispanic.

Indirect Identification of Hispanic/Latino 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/Latino ethnicity, some cancer registries have found it necessary to rely on indirect methods to populate this data element. No guidelines have been available to define the most valid approaches for indirect identification and thus lack of reliability in the resulting information across registry jurisdictions is also a concern. Registries often have significant numbers or proportions of Hispanic/Latino populations in their jurisdiction. They have needed to develop alternate approaches to enhance Hispanic/Latino identification that include reliance on death certificates, surname and maiden name matching algorithms, birth place, special studies, physician follow-up, and linkage with other data sources.

Based on a NAACCR survey of all registries and an empirical evaluation by representatives from states that produce cancer incidence data for the Hispanic/Latino population in their registry area, the initial guideline was that all registries follow the NAACCR Hispanic/Latino Identification Algorithm [1], and beginning in 2005 use NHIA v2.x. 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 1-3, and particularly the process in Figure 2; or 3) applying a computerized algorithm of these guidelines. For the third option, a SAS version is available for download from the NAACCR web site. The SAS version is bundled with the NAACCR Asian and Pacific Islander Identification Algorithm (NAPIIA), though it may be run independently. Registry staff also will need the 2000 Surname list from the U.S. Census. [5] The NHIA guideline was adapted from the Illinois State Cancer Registry Hispanic Algorithm. The Colorado Central Cancer Registry developed the original SAS version of the computerized NHIA algorithm. The New York State Cancer Registry staff modified the original program to run more efficiently, and staff of Information Management Services, Inc. implemented NHIA v2 and subsequent revisions.

The NAACCR Hispanic/Latino Identification Algorithm (NHIA), version 2.2.1

The NAACCR Hispanic/Latino Identification Algorithm, version 2.2.1 (NHIA v2.2.1) uses a combination of NAACCR variables to directly or indirectly classify cases as Hispanic/Latino for analytic purposes. It is possible to separate Hispanic/Latino ancestral subgroups (e.g., Mexican) when indirect assignment results from birthplace 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), Sex (item 220), and IHS Link (item 192). [6]

Only one race variable (Race 1, item 160) is considered in NHIA v2.2.1. This decision was based on the fact that only a very small percentage of cases have information on multi-race origins in their cancer registry. If this phenomenon changes in the future, in that there are

more cases reported with multiple races, then the decision should be revisited to expand the algorithm to capture information from NAACCR standard data items, Race 2 through Race 5.

As in previous versions, NHIA v2.2.1 can be applied to cases diagnosed from 1995 forward. Application of the method for the years prior to 1995 may be feasible, but each registry should determine its appropriateness for these earlier years.

Summary

Accurate execution of NHIA v2.2.1, as for previous versions, requires that the registry follow all NAACCR data standards, definitions, reporting rules, and codes. For example, following the NAACCR standard for maiden names, the field must be blank if maiden name is missing. If not, indirect assignment of ethnicity may not be correct.

A person is classified as **Hispanic** or **non-Hispanic** using NHIA v2.2.1 through either direct or indirect identification.

Direct Identification. Cases reported with one of the following codes on data element 190, Spanish/Hispanic Origin:

- 1 – Mexican (including Chicano);
- 2 - Puerto Rican;
- 3 - Cuban;
- 4 - South or Central American (except Brazil);
- 5 - Other specified Spanish/Hispanic origin (includes European);
- 8 – Dominican

Indirect Identification. Cases reported with one of the following codes on data element 190:

- 0 - non-Spanish/non-Hispanic;
- 7 - Spanish surname only;
- 9 - Unknown whether Spanish or not.

Direct and Indirect Identification. Cases reported with this code on data element 190:

- 6 - Spanish, NOS, Hispanic, NOS, Latino, NOS;

Persons are excluded from the indirect identification process if they are of Filipino, Hawaiian, or American Indian race (including Aleutian, Eskimo, and all indigenous populations of the Western hemisphere), or when they were born in certain countries (see Section 2.1 for specific list). These persons are classified as **non-Hispanic**.

Persons are also included as **Hispanic/Latino** when they are male cases with **heavily Hispanic/Latino** 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, occasionally Hispanic, or indeterminate** maiden names and **heavily Hispanic** last names.

If desired, following the specific options detailed below in Step 4 and based on local demographic information, a registry can exclude counties from the surname match portion of

the algorithm when the proportion of Hispanic/Latino residents in the 2000 U.S. Census population estimate of the county falls below 5%. [See Appendices A and B].

After applying NHIA v2.2.1, cases not classified as Hispanic/Latino are classified as **non-Hispanic**.

Detailed NHIA v2.2.1 Logic

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
8	Dominican (beginning with 2005 diagnoses)

For NAACCR standard data element 190, all cases reported by reporting facilities as Spanish/ Hispanic origin encompass codes 1, 2, 3, 4, 5, 6, and 8 (see table at left). This step represents the direct identification component of NHIA.

The indirect identification component involves cancer cases reported as Spanish/Hispanic origin data element codes 0, 6, 7 and 9 (see table below) for the NAACCR standard data element 190. The goal is to classify these cases as Hispanic/Latino, non-Hispanic, or a more specific Hispanic/Latino group based on an evaluation of the strength of the birthplace, race, and/or

Step1.2 Spanish/Hispanic Origin Data Element (NAACCR Data Element 190)	
Code	Category
0	Non-Hispanic
6	Spanish, NOS; Hispanic, NOS; Latino, NOS
7	Surname only
9	Unknown

or surname associations with Hispanic/Latino ethnicity status. While cases with code 6 are already known to be Hispanic, they may be assignable to a more specific category based on birthplace.

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 exclude such cases from the algorithm. 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.

Step 2. Filter Cases for Indirect Identification Based on Birthplace

2.1. Some cases are assigned to Hispanic/Latino ethnicity based on birthplace. Cases born in birthplaces associated with a high prevalence of Spanish surnames but a low probability of Hispanic/Latino ethnic status are excluded from the surname portion of the algorithm (see table). Anyone with a birthplace listed in the following table is coded to 0, non-Hispanic.

Step 2.1. Birthplaces Associated with Prevalence of Spanish Surnames but Low Probability of Hispanic/Latino Ethnicity	
FIPS Code	Birthplace
100, 102, 109	Atlantic/Caribbean area excluding Cuba , Dominican Republic, and Puerto Rico
110	Panama Canal
120-137	Pacific Area
331	Guyana
332	Suriname
333	French Guyana
341	Brazil
400-441; 445-499	Europe including Portugal (excluding Spain)
675	Philippines

2.2. In general, those cases born in birthplaces shown in the table for Step 2.2 have high probabilities of being Hispanic/Latino. 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/Latino ethnicity are classified **Hispanic/Latino** using NHIA v2.2.1.

Step 2.2. Birthplaces with High Probability of Hispanic Ethnicity					
Code	Birthplace	NHIA v2.1	Code	Birthplace	NHIA v2.1
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	8	321	Venezuela	4
250	Central America	4	345	Ecuador	4
251	Guatemala	4	351	Peru	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 Islands and Andorra).	5

Note that cases coded as 6 (Spanish, NOS; Hispanic, NOS; Latino, NOS) are excluded from steps 3-5. Steps 3-5 only apply to cases coded as 0, 7, or 9.

Step 3. Exclude Cases from Indirect Identification Based on Race

Cases reported with Race 1 (item 160) codes of 03 (American Indian, Aleutian or Eskimo, including all indigenous populations of the Western hemisphere), 06 (Filipino), 07 (Hawaiian), 96 (Asian NOS), and 97 (Pacific Islander NOS) are eliminated from indirect identification as these race groups often have Spanish surnames but not Hispanic ethnicity. This includes all cases with an IHS Link (item 192) value of 1, denoting a successful match to IHS.

Step 4. Filter Cases for Indirect Identification Based on County of Residence

At the discretion of a registry and upon their careful review of the validity of Hispanic/Latino origin assignment in counties with small numbers or small proportions of residents who self-identify as Hispanic/Latino in population counts from the U.S. Bureau of the Census, it is strongly encouraged that entire counties within a state be excluded from the surname matching portion of the algorithm. The Latino Research Group conducted an empirical analysis of CINA Deluxe data for 1995-2001, and based on indicators of sensitivity, specificity, and prevalence (i.e., positive predictive values), they recommend a threshold of 5%. In other words, for counties with fewer than 5% of the total population being of Hispanic/Latino ethnicity, it is strongly recommended that these counties be excluded from the surname match portion of the algorithm.

Thus, registries have the following options for counties in which less than 5% of the population is of Hispanic/Latino ethnicity:

1. Run the surname portion of the algorithm only on cases reported on data element 190, as Spanish surname only or as unknown whether Spanish (item 190 – codes 7 or 9). [See Appendix A].
2. Run the surname portion of the algorithm only on cases with a code of 7 on data element 190 (to verify that the surname is on the list of allowable Hispanic surnames) AND convert all cases with a code of 9 (unknown if Hispanic) to a code of 0 (Not Hispanic).

With both options, the surname portion will not be run on cases coded as 0, non-Hispanic.

Note that when choosing either of these options, it is possible for a person to be classified as Hispanic for one tumor and non-Hispanic for another tumor, if he/she changed counties between diagnoses.

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

In step 5, the Last and Maiden Surnames are categorized according to the 2000 Census Bureau Surname List. This list is based on the complete count of the 2000 census and contains all surnames occurring at least 50 times with percentages for detailed race/ethnicity groups including Hispanic. Names with Hispanic prevalence over 75 percent are considered “heavily” Hispanic, >50-75 percent are considered “generally” Hispanic, >25-50 percent are considered “moderately” Hispanic, >5-25 percent are considered “occasionally” Hispanic, and 5 percent or below are considered “rarely” Hispanic. Names not on the list are also considered rarely Hispanic.

For females, indirect identification is based on both maiden name and last name:

- Female cases with heavily Hispanic maiden names are classified as Hispanic (code 7).
- Female cases with rarely Hispanic maiden are classified as non-Hispanic (code 0).
- Remaining female cases are classified as Hispanic (code 7) if their last names are heavily Hispanic, otherwise they are classified as non-Hispanic (code 0).

Step 6. Save the results of NHIA v2.2.1 as a separate data element.

Step 6. NHIA v2.2.1 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 v2.2.1 surname match only
8	Dominican, by birthplace or other specific identifier (became a standard with diagnoses 01/01/2005)

The results of NHIA v2.2.1 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 at left. The one exception is that no missing codes will be allowed, because at the conclusion of step 6 of NHIA v2.2.1, if a case has not been identified as Hispanic/Latino, it will be coded to 0, non-Hispanic. The NHIA v2.2.1 variable is placed in column number 231 in version 12 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 race, Hispanic ethnicity, and any combinations thereof.
- 4.** Run the algorithm for all cases with data element #190, Spanish/Hispanic Origin coded to either a 0 (non-Spanish; non-Hispanic), a 6 (Spanish, NOS; Hispanic, NOS; Latino, NOS), 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, this information will need to be supplied to NAACCR with a file submitted in response to a Call for Data. This procedure may be applied to all cases and not just be limited to cases in counties that do not meet the 5% threshold of the total population being of Hispanic/Latino ethnicity.
- 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.

References

1. 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.
2. Howe HL, Carozza S, O'Malley C, Dolecek TA, Finch JL, Kohler B, Wet D, Liu L, Schymura MJ, Williams M, Abe T, Agovino P, Chen VW, Firth R, Harkins D, Hotes, J, Lake A, Roney D, Suarez L (eds). *Cancer in U.S. Hispanics/Latinos, 1995-2000*. Springfield, IL: North American Association of Central Cancer Registries, December 2003.
3. Howe HL. *Evaluation of NHIA Submissions for 1997-2001*. Springfield, IL; North American Association of Central Cancer Registries, January 2005.
4. Ellison JH, Wu XC, Howe HL, McLaughlin CC, Lake A, Firth R, Sullivan SK, Roney D, Cormier M, Leonfellner S, Kosary C (eds). *Cancer in North America, 1998-2002. Volume Four: Cancer Incidence in U.S. Hispanic/Latino Populations*. Springfield, IL: North American Association of Central Cancer Registries, April 2005.
5. U.S. Census Bureau. Genealogy Data: Frequently Occurring Surnames from Census 2000. On-line: <http://www.census.gov/genealogy/www/data/2000surnames/index.html>. Accessed September 12, 2011.
6. Thornton ML, O'Connor L, editors. *Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Fourteenth Edition, Version 12*. Springfield, IL: North American Association of Central Cancer Registries, February 2009.

Appendix A. Sensitivity and Specificity of Heavily Hispanic Surnames based on the 1990 Census Spanish Origin Research file

		Heavily Hispanic Surname		Total
		+	-	
Self-reported Ethnicity Hispanic	+	135,131	25,040	160,171
	-	7,670	879,638	887,308
	Total	142,801	904,678	1,047,479

Source: Used with permission of Carin Perkins of the Minnesota Cancer Surveillance System

1,047,479 householders had a surname that was given by at least one self-reported Hispanic. Of these, 160,171 (15.3%) were Hispanic by self-report. Of the 25,276 Hispanic Surnames on the file, 12,215 were “heavily” Hispanic.

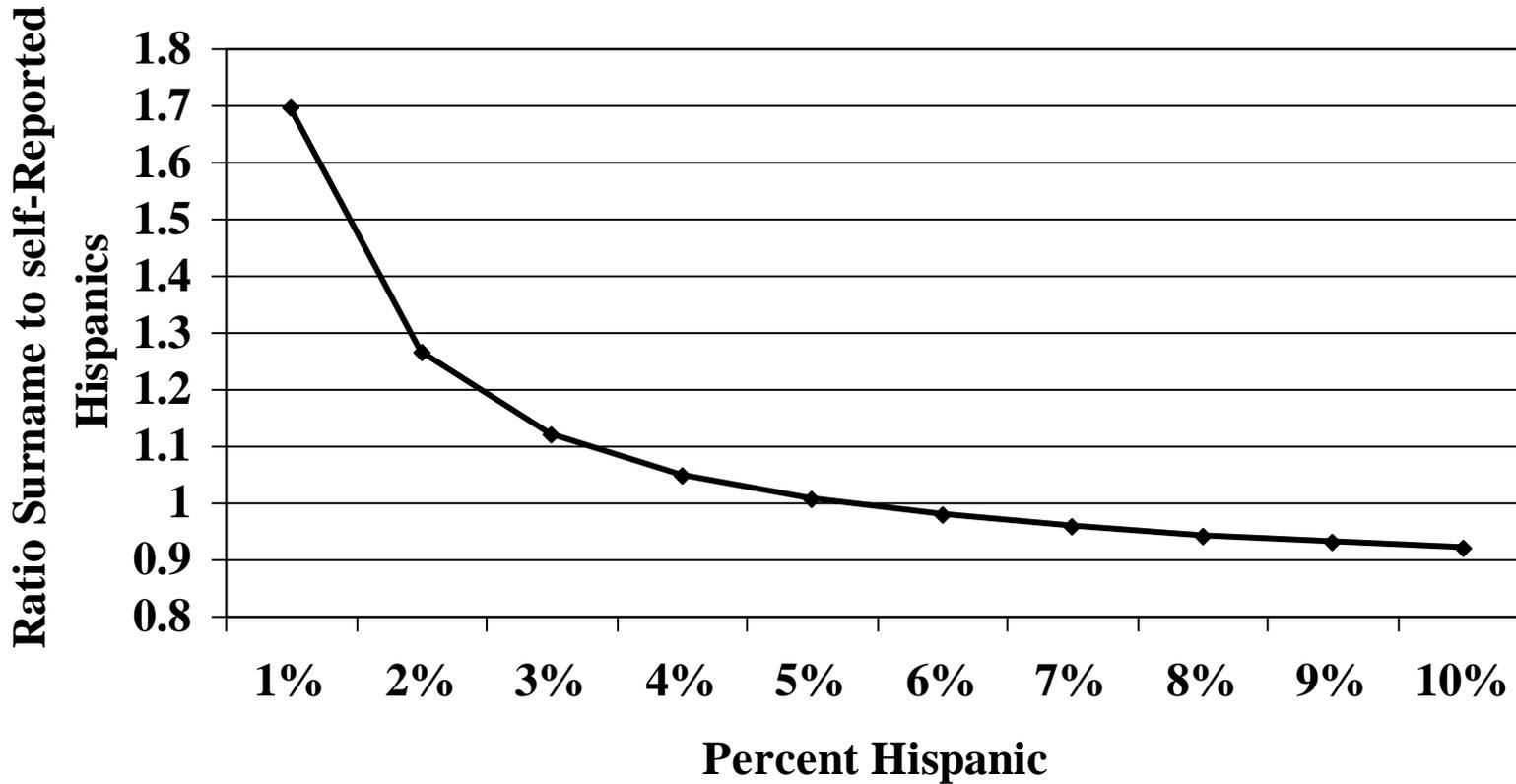
Sensitivity = proportion of self-reported Hispanics with heavily Hispanic surname
 $= 135,131 / 160,171 = 84.37\%$

Specificity = proportion of non-Hispanics who don't have heavily Hispanic surnames
 $= 879,638 / 887,308 = 99.14\%$

Predictive value positive = proportion of heavily Hispanic surnames who are Hispanic
 $= 135,131 / 142,801 = 94.6\%$

Ratio of surname Hispanics to self-reported = $142,801 / 160,171 = 0.8915$

Appendix B. Inflation of Hispanic cases using 1990 Census heavily Hispanic surnames to identify Hispanics as a function of the proportion of Hispanics in the population



Sensitivity = 0.8437; specificity = 0.9914.

Source: Used with permission of Carin Perkins of the Minnesota Cancer Surveillance System