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

Revised September 21, 2005

Editors:

NAACCR Latino Research Work Group

Chaired by
Holly L. Howe, PhD
Executive Director, NAACCR

As of September 21, 2005

Suggested Citation:
NAACCR Latino Research Work Group. *NAACCR Guideline for Enhancing Hispanic/Latino Identification: Revised NAACCR Hispanic/Latino Identification Algorithm [NHIA v2]*. Springfield (IL): North American Association of Central Cancer Registries. September 2005.

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 Latino Research Group

Holly L. Howe, PhD, Chair
NAACCR
2121 W White Oaks Dr
Springfield, IL 62704
(217) 698-0800
hhowe@naaccr.org

Pamela Agovino, MPH
NJ Dept. of Health & Senior Services
Cancer Epidemiology Services
PO Box 369
Trenton, NJ 08625-0369
(609) 588-3500
Pamela.Agovino@doh.state.nj.us

Cheryll Cardinez
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention
Davidson Buidling Rm 3246
2858 Woodcock Blvd
Atlanta GA 30341-3724
zzg3@cdc.gov

Susan Carozza, PhD
Department of Epidemiology & Biostatistics
School of Rural Public Health
TAMUS Health Science Center
3000 Briarcrest Dr., Suite 310
Bryan, TX 77802
(979) 862-8168
scarozza@srph.tamhsc.edu

Vivien W. Chen, PhD
Louisiana Tumor Registry
Louisiana State University Health Sciences Center
1600 Canal St, Suite 900A
New Orleans, LA 70112
(504) 568-6047
vchen@lsuhsc.edu

Jack Finch, MS
Colorado Central Cancer Registry
Colorado Department of Public Health & Environment
4300 Cherry Creek Dr S
Denver, CO 80222-1530
(303) 692-2544
jack.finch@state.co.us

Mei-chin Hseih
Louisiana Tumor Registry
Louisiana State University Health Sciences Center
1600 Canal St, Suite 900A
New Orleans, LA 70112
(504) 568-6047

Betsy Kohler, MPH, CTR
NJ Dept. of Health & Senior Services
Cancer Epidemiology Services
PO Box 369
Trenton, NJ 08625-0369
(609) 588-3500
betsy.kohler@doh.state.nj.us

Andrew Lake
IMS
12501 Prosperity Drive
Suite 200
Silver Spring, MD 20904
(301) 680-9770
lakea@imsweb.com

Lihua Liu, PhD
Los Angeles Cancer Surveillance Program
Keck School of Medicine
University of Southern California
1540 Alcazar St, CHP 204
Los Angeles, CA 90033
(323) 442-2300
lihualiu@usc.edu

Barry Miller, DrPH
National Cancer Institute
SEER Program, SRP, DCCPS
6116 Executive Blvd, MSC 8315
Suite 504
Bethesda, MD 20892-8316
(301) 402-4248
millerb@mail.nih.gov

Cynthia O'Malley, PhD
Northern California Cancer Center
2201 Walnut Avenue, Suite 300
Fremont, CA 94538
(510) 608-5034
comalley@nccc.org

Carin Perkins, PhD
Minnesota Cancer Surveillance System
Minnesota Department of Health
717 Delaware SE
P. O. Box 9441
Minneapolis, MN 55440-9441
(612) 676-5657
carin.perkins@health.state.mn.us

David Roney
IMS
12501 Prosperity Drive
Suite 200
Silver Spring, MD 20904
(301) 680-9770
roneyd@imsweb.com

Maria Schymura, PhD
NY State Cancer Registry
Corning Tower Room 536
Empire State Plaza
Albany, NY 12237-0679
(518) 474-2255
mjs08@health.state.ny.us

Melanie Williams, PhD
Texas Cancer Registry
Cancer Epidemiology and Surveillance Branch
Texas Department of State Health Services
1100 West 49th Street
Austin TX 78756-3199
512.458.7523
Melanie.Williams@tdh.state.tx.us

Table of Contents

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. *[2]*

Cancer Incidence in U.S. Hispanics/Latinos, 1995-2000 (CIUSHL) was published in December 2003. *[3]* 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. *[4]* 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]. [5]*

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 to the original NHIA are described below as the NAACCR Hispanic/Latino Identification Algorithm version 2 [NHIA v2], released in 2005.

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 true 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. 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 (a SAS version is available for download from the NAACCR web site). Registry staff also will need the 1990 Hispanic Surname list from the U.S. Census. *[6]* 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. did the 2005 modifications for NHIA v2.

The NAACCR Hispanic/Latino Identification Algorithm, version 2 (NHIA v2)

The NAACCR Hispanic/Latino Identification Algorithm, version 2 (NHIA v2) 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), and Sex (item 220). *[7]*

In 2005, only one race variable is considered in NHIA v2, Race 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.

Similar to NHIA version 1, NHIA v2 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. *[2]*

Summary

Accurate execution of NHIA v2, as for NHIA v1, 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 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. *[7]*
8 – Dominican

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

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

Persons are excluded from the indirect identification process if they are of Filipino, Native American (including indigenous tribes of Latin America determined using the 2004 SEER Program Manual on Coding Race), or Hawaiian race 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, cases not classified as Hispanic/Latino are classified as **non-Hispanic**.

Detailed NHIA v2 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)
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. 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 v2).

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/Latino or non-Hispanic based on an evaluation of the strength of the birthplace, race, and/or surname associations with Hispanic/Latino ethnicity status.

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.

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 of Step 2.1). 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.

Step 2.2. Birthplaces with High Probability of Hispanic Ethnicity

FIPS Code	Birthplace	NHIA v2	FIPS Code	Birthplace	NHIA v2
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
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 3. 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 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, entire counties within a state

may be excluded from the surname matching portion of the algorithm. The Latino Research Group has 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, these counties **may** 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).

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

Of course, these two options are available for registries that choose to exercise the 5% threshold. These options are just that, options. A registry can choose to apply NHIA v2 to all cases regardless of the population density of persons of Latino descent.

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

In step 5, the Last and Maiden Surnames are categorized 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 and assign probability codes to registry cases. *[6]* 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 5.

Step 5. 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

For males, cases with last names coded to heavily Hispanic (101, 102, 105, 110, 115 or 125) are classified as Hispanic (code 7). The remaining male cases are classified as non-Hispanic (code 0).

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

- Female cases with maiden names coded to heavily Hispanic (101, 102, 105, 110, 115 or 125) are classified as Hispanic (code 7).
- Female cases with maiden names coded to rarely Hispanic (5001, 5005, 5010, 5025, 5100, or 5500) are classified as non-Hispanic (code 0).

The remaining female cases [i.e., those whose maiden name is

- missing (9000),

- indeterminate (6000), or
- classified as generally, moderately or occasionally Hispanic (201, 202, 205, 210, 215, 225, 301, 302, 305, 310, 315, 325, 405, 410, 415, or 425)]

are classified as Hispanic if their last names are coded to heavily Hispanic (101, 102, 105, 110, 115 or 125); otherwise they are classified as non-Hispanic (code 0).

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

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

The results of NHIA v2 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 v2, if a case has not been identified as Hispanic/Latino, it will be coded to 0, non-Hispanic. The NHIA v2 variable is placed in column number 231 in both NAACCR Data Exchange Layouts, versions 10.2 and 11.

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, 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.

NHIA v2 Diagrams

Notation for Diagrams

Unified Modeling Language (UML) diagramming instruments were used to describe the NAACCR Hispanic identification process. UML, a standard modeling language, represents a collection of the best engineering practices that have proven to be successful in modeling large and complex systems and processes. Two diagrams (Figures 1 and 2) depict the NAACCR Hispanic/Latino identification v2 process. The diagram legend presented in Figure 3 describes elements (building blocks) of diagrams and is intended to help readers navigate the diagrams and interpret their meanings.

Elements used in UML class and activity diagrams of Hispanic identification process include (see Figure 3 for a pictorial representation):

- **Class.** Represents a tangible thing, for example – a data item.
- **Use Case.** Represents a process in the form of narrative/description (scenario).
- **Activity node.** Represents a business process, operation, or activity. The activity node is shown as a shape with a straight top and bottom, and with convex arcs on the two sides. Activity nodes are shaded blue. Examples of activity nodes in Figure 5 are *Evaluate Data Item 190 code values and Assign Surname Codes from the Spanish Surnames List.*
- **Note.** Used for descriptive text. A note is shown as a dog-eared rectangular shape with its upper-right corner bent.
- **Object.** Represents inputs and outputs for activity nodes. An object is shown as a rectangle. Examples of objects in Figure 5 include *Spanish Surnames List and 3: Set of Cases for Indirect Identification.*
- **Object Flow (Input/Output).** Input/Output connection between an object and an activity node (*object flow*) is shown as an arrow.
- **Synchronization bar (horizontal or vertical).** Used to depict parallel processing. A synchronization bar is shown as a bold black line.
- **Transition.** Transition between activities is shown as a solid black arrow.

Diagrams Descriptions

The diagram in Figure 1 (Elements of NAACCR Hispanic Identification Process) depicts data items relevant for Hispanic/Latino identification in a central cancer registry. A class (rectangular shape), color-coded green, represents each data item. Lists of data items codes are represented with classes (rectangular shapes), color-coded yellow. Information external for the NAACCR case record layout includes Spanish Surname List, which is color-coded white. Hispanic identification process is depicted with a blue oval shape. Dashed arrows on this diagram indicate "dependency" relationships: Hispanic identification process depends on getting information from data items 190, 2230, 2390, 220, 250, 90 and 160; a new data item (similar to data element 190) depends on results of Hispanic identification process to assign proper code values to cancer cases.

The diagram in Figure 2 (Process Map: NAACCR Guideline for Hispanic Identification in a Central Cancer Registry) represents the algorithm for Hispanic identification in a central cancer registry. The process is portrayed with activities (blue shapes) and objects (rectangular shapes) that represent inputs and outputs (products) of activities. Hispanic identification includes consecutively implemented filtering steps that stratify initial set of cancer cases based on filtering criteria. Each set of cases is a result of accumulated decisions made on all the previous filtering steps. Such accumulation is reflected inside of the objects (rectangular shapes) that represent sets of cases. For example, Set of Cases for Indirect Identification (marked number 32 at the top right corner of the box) contains cases with Data Item 190 code values 7, 9, and 0 (optional) – according with step 1, as indicated with "1A" and "1B"; plus these

cases were filtered out with the Birthplace criteria (according with step 2, as indicated with "2C") and with the Race criteria (according with step 3, as indicated with "3B"). Additionally, the number 32 points out that this set of cases is produced at the step 3.

There are two possible outcomes for each cancer case from NHIA v2: Hispanic (sets of cases/objects color-coded pink) or non-Hispanic (sets of cases color-coded light brown). Intermediate sets of cases ("in-process" sets/objects) are color-coded with green. Objects with supporting information, like code values for data items, are color-coded yellow. A new Data Item (190-like) for cases identified as non-Hispanic is updated with a code 0 (zero), and for cases identified as Hispanic, with codes 1through 8.

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. Minutes from the NAACCR Collaborative Research Work Group (CRWG) Expert Panel on Hispanic Identification, November 4, 2003.

3. 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.

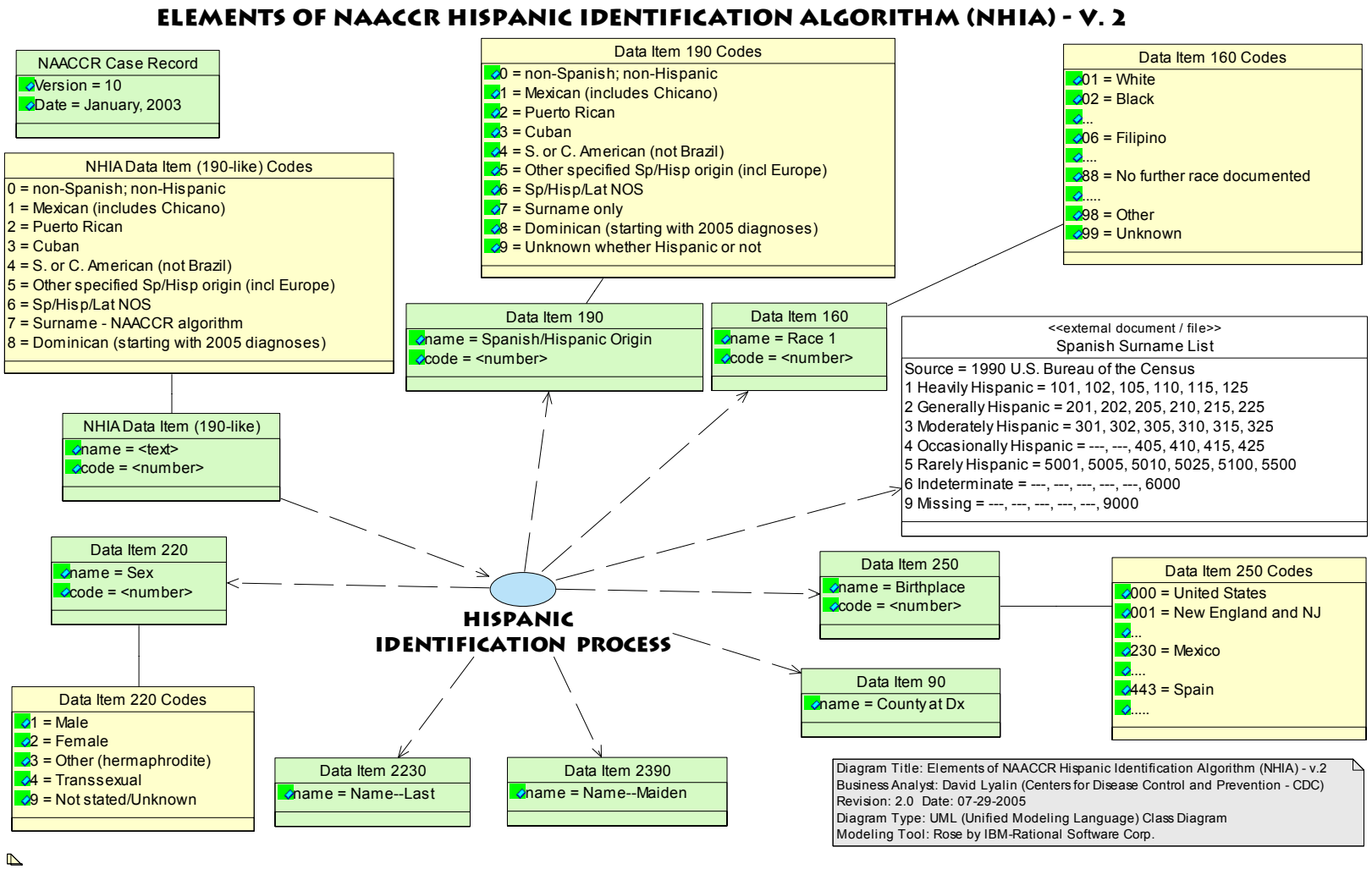
4. Howe HL. Evaluation of NHIA Submissions for 1997-2001. Springfield, IL; North American Association of Central Cancer Registries, January 2005.

5. 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, Inc. April 2005.

6. Word DL, Perkins RC, Jr. *Building a Spanish Surname List for the 1990's* B *A New Approach to An Old Problem.* Population Division Working Paper No. 13. Washington DC: U.S. Bureau of the Census. March 1996. [http://www.census.gov/population/documentation/twpno13.pdf, accessed February 16, 2003].

7. Havener L, Hultstrom D, editors. Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Tenth Edition, Version 11. Springfield, IL: North American Association of Central Cancer Registries, October 2004.

Figure 1


ELEMENTS OF NAACCR HISPANIC IDENTIFICATION ALGORITHM (NHIA) - V. 2
NAACCR Case Record
Version = 10
Date = January, 2003
NHIA Data Item (190-like) Codes
0 = non-Spanish; non-Hispanic
1 = Mexican (includes Chicano)
2 = Puerto Rican
3 = Cuban
4 = S. or C. American (not Brazil)
5 = Other specified Sp/Hisp origin (incl Europe)
6 = Sp/Hisp/Lat NOS
7 = Surname - NAACCR algorithm
8 = Dominican (starting with 2005 diagnoses)
Data Item 190 Codes
0 = non-Spanish; non-Hispanic
1 = Mexican (includes Chicano)
2 = Puerto Rican
3 = Cuban
4 = S. or C. American (not Brazil)
5 = Other specified Sp/Hisp origin (incl Europe)
6 = Sp/Hisp/Lat NOS
7 = Surname only
8 = Dominican (starting with 2005 diagnoses)
9 = Unknown whether Hispanic or not
Data Item 160 Codes
01 = White
02 = Black
...
06 = Filipino
....
88 = No further race documented
.....
98 = Other
99 = Unknown
Data Item 190
name = Spanish/Hispanic Origin
code = <number>
Data Item 160
name = Race 1
code = <number>
<<external document / file>>
Spanish Surname List
Source = 1990 U.S. Bureau of the Census
1 Heavily Hispanic = 101, 102, 105, 110, 115, 125
2 Generally Hispanic = 201, 202, 205, 210, 215, 225
3 Moderately Hispanic = 301, 302, 305, 310, 315, 325
4 Occasionally Hispanic = ---, ---, 405, 410, 415, 425
5 Rarely Hispanic = 5001, 5005, 5010, 5025, 5100, 5500
6 Indeterminate = ---, ---, ---, ---, ---, 6000
9 Missing = ---, ---, ---, ---, ---, 9000
NHIA Data Item (190-like)
name = <text>
code = <number>
Data Item 220
name = Sex
code = <number>
Data Item 250
name = Birthplace
code = <number>
Data Item 250 Codes
000 = United States
001 = New England and NJ
...
230 = Mexico
....
443 = Spain
.....
HISPANIC IDENTIFICATION PROCESS
Data Item 90
name = County at Dx
Data Item 220 Codes
1 = Male
2 = Female
3 = Other (hermaphrodite)
4 = Transsexual
9 = Not stated/Unknown
Data Item 2230
name = Name--Last
Data Item 2390
name = Name--Maiden
Diagram Title: Elements of NAACCR Hispanic Identification Algorithm (NHIA) - v.2
Business Analyst: David Lyalin (Centers for Disease Control and Prevention - CDC)
Revision: 2.0 Date: 07-29-2005
Diagram Type: UML (Unified Modeling Language) Class Diagram
Modeling Tool: Rose by IBM-Rational Software Corp.

PROCESS MAP: NAACCR HISPANIC IDENTIFICATION ALGORITHM (NHIA) - V. 2

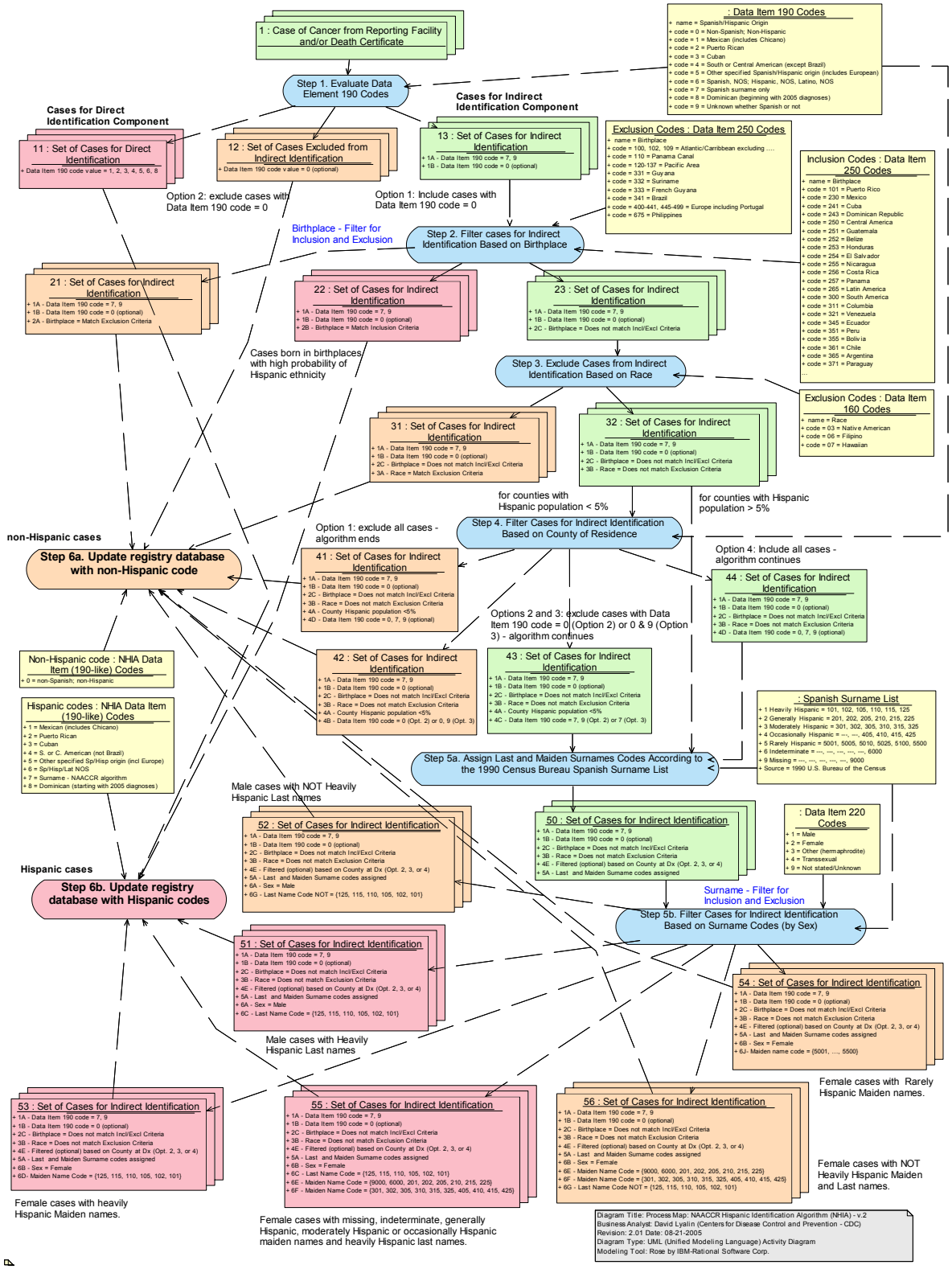
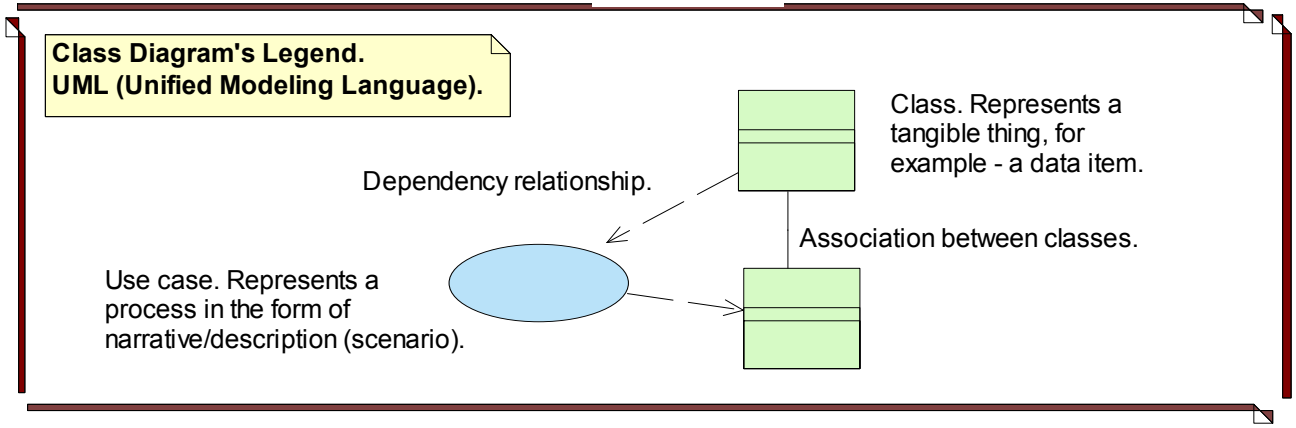
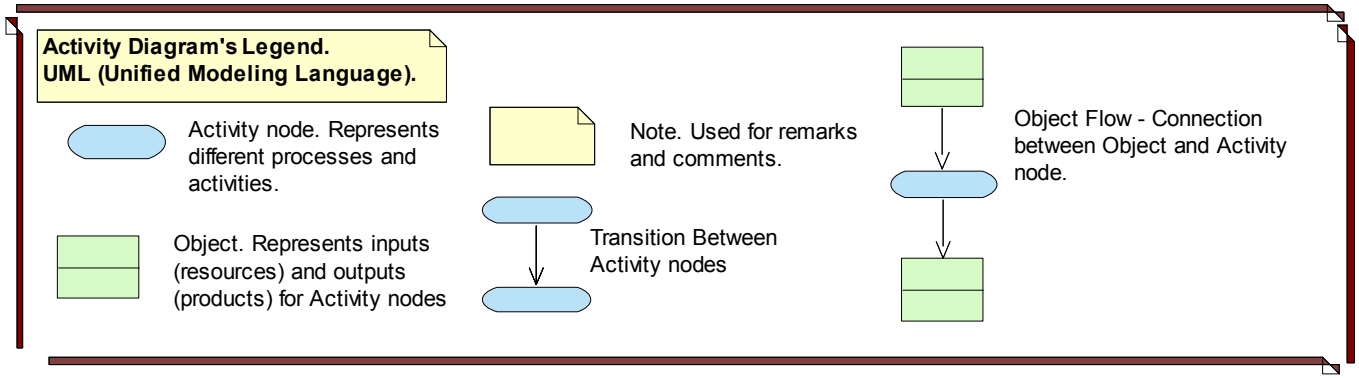


Figure 3.


Class Diagram's Legend.
UML (Unified Modeling Language).
Class. Represents a tangible thing, for example - a data item.
Dependency relationship.
Association between classes.
Use case. Represents a process in the form of narrative/description (scenario).


Activity Diagram's Legend.
UML (Unified Modeling Language).
Activity node. Represents different processes and activities.
Note. Used for remarks and comments.
Object Flow - Connection between Object and Activity node.
Object. Represents inputs (resources) and outputs (products) for Activity nodes
Transition Between Activity nodes

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

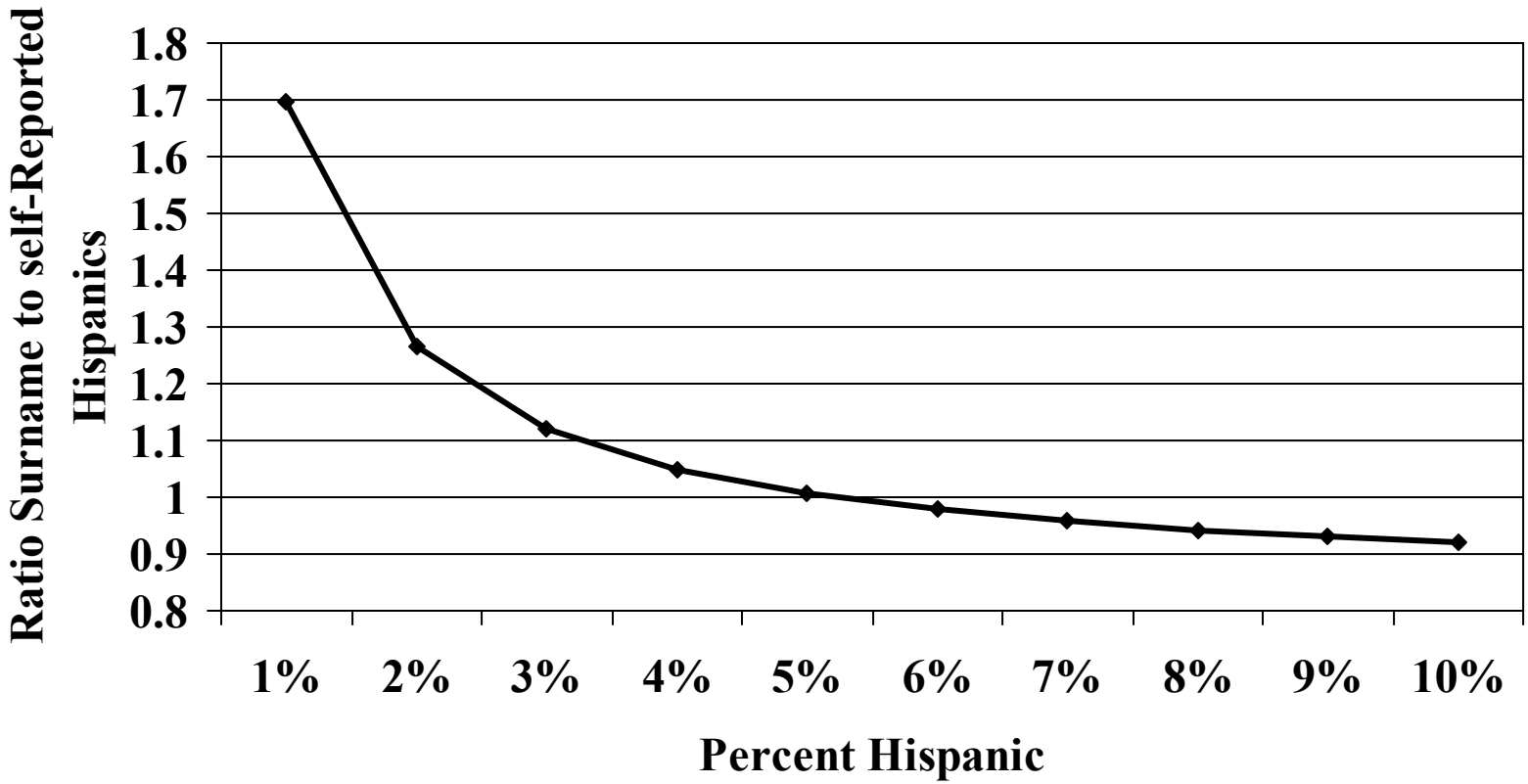
		Heavily Hispanic Surname		
		+	-	Total
Self-reported Ethnicity Hispanic	+	135,131	25,040	160,171
	-	7,670	879,638	887,308
	Total	142,801	904,678	1,047,479

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