

Hispanic Identification
in the
Illinois State Cancer Registry

by

Therese A. Dolecek, Ph.D.
Holly L. Howe, Ph.D.

Division of Epidemiologic Studies

A Publication of the
Illinois Department of Public Health
Division of Epidemiologic Studies
Illinois State Cancer Registry
Springfield, IL 67261

June 1998

Acknowledgments

The authors express their gratitude to the staff of the Illinois State Cancer Registry and to all of the staff members at reporting facilities for providing the data that made the study of Hispanic identification in the Illinois State Cancer Registry possible.

Special thanks to Mr. David Word with the United States Bureau of the Census and senior author of the working paper, *Building a Spanish Surname List for the 1990's A New Approach to An Old Problem*, for providing guidance in the application of the approach to cancer registry data in Illinois.

Suggested citation

Dolecek TA, Howe HL. Hispanic Identification in the Illinois State Cancer Registry. Epidemiologic Report Series 98:2. Springfield, Ill.: Illinois Department of Public Health, June 1998.

Copyright information

All material in this report is in the public domain and may be reproduced or copied without permission; citation as to source, however, is appreciated.

Table of Contents

Executive Summary	1
Introduction	3
Objectives	3
Methods	5
Information Sources for Hispanic Identification	5
ISCR Hispanic Origin Data Element	5
Birthplace and Race	5
Surnames	6
Application in Illinois	7
Procedures	7
Results and Discussion	13
ISCR Hispanic Origin Data Element	13
ISCR Hispanic Ancestral Subgroup Distribution Compared with SEER 1988 to1992	13
ISCR Hispanic Identification for Agreement Compared with Illinois Death Certificates	15
Birthplace	15
Comparison of ISCR Surnames with Census Spanish Surname Study	16
ISCR Surnames by Collapsed Census Spanish Surname Study Categories	17
Surnames for Cancer Cases Born in Hispanic Non-U.S. Birthplaces	18
American Indian, Filipino and Hawaiian Cancer Case Last Names	18
Effects of Surname Adjustment on Hispanic Non-U.S. Birthplace As Hispanic Identifier	18
Hispanic Identification	19
Hispanic Cancer Incidence Rates	21
Conclusions	25
Limitations	27
References	29
Tables	31
Appendix A. Additional Tables	71
Appendix B. Alphabetical and Categorical Listing of United States Bureau of the Census Heavily Hispanic Last Names on the ISCR Database 1986 to1994	86

Executive Summary

A comprehensive study of Hispanic identification in the Illinois State Cancer Registry (ISCR) was conducted for the purpose of establishing a policy that defines the reporting of Hispanic cancer incidence in Illinois. In 1989, an Hispanic origin data element was introduced on the ISCR incidence report form as a requirement for data acquisition by cancer registrars in Illinois. It was found that cancer registrars were able to identify approximately 75 percent to 80 percent of Hispanic cancer cases using resources available to them. Indirect identifiers including birthplace and surname were examined to determine if development of an algorithm that used both direct and indirect information on Hispanics could improve case ascertainment on ISCR.

Birthplaces not in the United States (U.S.) that suggest probable Hispanic ethnic status include Central America (Guatemala, Belize, Honduras, El Salvador, Nicaragua, Costa Rica and Panama), Cuba, Mexico, Puerto Rico, South America (Colombia, Venezuela, Ecuador, Peru, Bolivia, Chile, Argentina, Paraguay and Uruguay) and Spain (Canary Islands, Balearic Islands and Andorra). An evaluation of cases born in Hispanic non-U.S. birthplaces indicated that the use of birthplace information could enhance Hispanic casefinding on ISCR.

In addition, Spanish surnames are known to be predictors of Hispanic status. In March 1996, a Spanish surname list developed through special study of 1990 census surnames and Hispanic origin responses, became available to facilitate Hispanic identification in the U.S. Evaluation of the census Spanish surname approach indicated that it was applicable to the Hispanic population in Illinois and was suitable for indirect Hispanic identification of cases on the ISCR database.

It was determined that a combination of direct and indirect Hispanic identification was necessary to produce the most complete ascertainment of Hispanic cancer cases. An algorithm was created using both direct (ISCR Hispanic origin) and indirect (Hispanic non-U.S. birthplace and heavily Hispanic surname as classified according to the census bureau Spanish surname study) identifiers that maximized the probability of correct Hispanic identification of ISCR cases.

Hispanic cancer incidence will be reported using the established algorithm for reports and special studies in the Division of Epidemiologic Studies. More valid and reliable cancer incidence data on the Hispanic population residing in Illinois should guide future cancer control and prevention programs for this rapidly growing ethnic group.

Introduction

In recent decades, the Hispanic population has been the most rapidly growing race/ethnic group within the United States (U.S.) increasing, by some estimates, at a rate five times that of the rest of the U.S. population.¹ Due to high birth rates and immigration patterns among the group, such growth is projected to continue well into the next century. By the year 2000, Hispanics may become the largest minority group in the nation.²

A population presence of this nature necessitates targeted health care planning at federal, state and local levels. Such activities require valid and reliable data on the respective population to form a basis for accurate needs assessments and subsequent program planning, implementation and surveillance. However, program development for Hispanics has been impeded due to a lack of health-related data on the group. Only recently have changes in the organized U.S. data collection systems specifically targeted the Hispanic population.

This problem is especially important for Illinois because it is among the 11 states where the country's Hispanic population is concentrated. Illinois' Hispanic population numbered approximately 910,000 in 1990, or about 7.9 percent of the state's total population.³ In 1980, the number of Hispanics in Illinois was approximately 636,000, or 5.6 percent of the total state population.⁴ Intercensal estimates for 1996 show Hispanics to number over 1.1 million, now about 9.6 percent of Illinois' total population.³ The dynamics observed for the nation's Hispanic population also are apparent for Hispanics residing in Illinois.

Therefore, the purpose of this study was to conduct a comprehensive evaluation of current Hispanic identification in the Illinois State Cancer Registry. In addition, the use of algorithms to validate or to enhance identification using indirect sources was studied with the intention of producing the most accurate information possible on cancer incidence among Illinois' Hispanic population to guide future cancer control and prevention programs.

Objectives

1. To assess Hispanic identification across the ISCR database from 1986 to 1994
2. To determine the applicability of a U.S. Bureau of the Census Spanish surname product, made available in March 1996, to Hispanic identification needs by ISCR
3. To establish an effective methodology for identifying Hispanic cancer cases on the ISCR database using all relevant information available for the study of cancer incidence trends among the ethnic group

Methods

Information Sources for Hispanic Identification

ISCR and other cancer registries have used several sources of information to study Hispanic identification on their databases.⁵ Specific Hispanic origin data elements, birthplace, race and surnames are variables found to be important for direct or indirect identification. The ISCR database contains data in varying degrees of completeness for all of these Hispanic information sources.

ISCR Hispanic Origin Data Element

Table 1 shows a historical presentation of the Hispanic origin data elements from inception of ISCR in 1985 to the present. Initially, Hispanic origin was not collected using a separate element on the incidence report form. Rather, cancer registrars documented information on Hispanic status using medical record information and birthplace data under the race/ethnicity element. A revision of the ISCR incidence report form in May 1989 included a separate Hispanic origin item that allowed a yes or no response for Hispanic status and the ability to specify Hispanic subgroup if such information were available to the registrar. In January 1993, another revision incorporating the North American Association of Central Cancer Registries (NAACCR) format and codes specified Hispanic subgroups to include Mexican, Puerto Rican, Cuban, South or Central American, other Spanish (includes European), Spanish no other specification (NOS) or unknown.⁶ A subsequent revision in May 1995 further refined the format by the addition of a Spanish surname only category. The last two revisions closely resemble the Hispanic origin item included on the 1990 census form.

Birthplace and Race

Because of the high migration rates of Hispanic persons to the U.S., information on birthplace is useful as an indirect identifier of Hispanic status. Birthplaces that suggest probable Hispanic ethnic status include Central America (Guatemala, Belize, Honduras, El Salvador, Nicaragua, Costa Rica and Panama), Cuba, Mexico, Puerto Rico, South America (Colombia, Venezuela, Ecuador, Peru, Bolivia, Chile, Argentina, Paraguay and Uruguay) and Spain (Canary Islands, Balearic Islands and Andorra). Conversely, birthplace also may serve as a means to refine methodologies by avoiding erroneous identification of individuals where Spanish surnames may prevail but classification as Hispanic is incorrect. These birthplaces include Asia (especially the Philippines), the Atlantic/Caribbean area (excluding Cuba and Puerto Rico), Europe (excluding Spain), the Pacific area, the Panama canal zone and South America (including Brazil, Guyana, Suriname and French Guiana). These birthplaces are shown with geocodes used by ISCR in Table 2.

Among certain race groups, Spanish names are prevalent but these persons are usually not Hispanic and should be excluded during identification procedures. Such race groups include American Indian, Filipino and Hawaiian. Data on birthplace and race have been collected by ISCR since inception of the registry.

Surnames

Spanish surname lists have been available from the U.S. Bureau of the Census for 1950, 1960, 1970 and 1980. These resources have been used to identify Hispanic individuals for a variety of purposes including identification on cancer registry databases. The lists have specified a surname as Spanish and names not listed have been assumed to be non-Spanish.

In March 1996, a Spanish surname product became available using data from the 1990 census. Unlike previous products, this version offers a more refined approach to Hispanic identification using surnames.

The development of this product made use of 5,609,592 census records with name appended for the purpose of determining undercount. These records included both a valid surname and a response to the 1990 Hispanic origin question allowing the cross-tabulation of Spanish origin question responses by surname and the ability to group surnames based on frequency of responses.

From these records, surnames were identified for 1,868,781 "householders" to mitigate clustering effects because most persons within a household also have the same surname and ethnicity. Therefore, the generalizability of results from this census bureau study is greatest for male householders and females who have never married.

A classification scheme was created from the database used for the 1990 census Spanish surname study. Using both the relationship of the surname to the Hispanic origin response and the frequency of occurrence of the name in the sample, these 28 Spanish surname categories were established.

Heavily Hispanic	Generally Hispanic	Moderately Hispanic	Occasionally Hispanic	Rarely Hispanic
0101	0201	0301	0405	5001
0102	0202	0302	0410	5005
0105	0205	0305	0415	5010
0110	0210	0310	0425	5025
0115	0215	0315		5100
0125	0225	0325		5500

The first two characters denote Hispanic classification (01 heavily Hispanic, 02 generally Hispanic, 03 moderately Hispanic, 04 occasionally Hispanic and 50 rarely Hispanic). The criterion for placing a surname within each category depended upon the proportion of householders with the surname responding affirmatively to the Hispanic origin item on the 1990 census form. The criterion for each of the five Spanish surname categories was as follows.

Spanish Surname Classification	Proportion of Householders Who Are Hispanic
01 Heavily Hispanic	Over 75 Percent
02 Generally Hispanic	50 Percent < x ≤ 75 Percent
03 Moderately Hispanic	25 Percent < x ≤ 50 Percent
04 Occasionally Hispanic	5 Percent < x ≤ 25 Percent
50 Rarely Hispanic	5 Percent ≥ x

The third and fourth characters represent the frequency with which a name occurred in the study sample. (For example, category 0125 includes surnames that are heavily Hispanic with at least 25 observations.)

In addition, other characteristics indicating strengths or weaknesses of the association between surname and Hispanic status, including Spanish state, orthographic structure of the name (Buechley rules) and status relative to the 1980 Spanish surname list, were examined when the schematic coding system was developed. The approach is described in more detail elsewhere.⁷

The ISCR database contains complete data on last names of all reported cancer cases. Maiden names also are documented by cancer registrars but these data are less complete. Marital status is not included on the ISCR incidence report form.

Application in Illinois

An Hispanic identification evaluation study was conducted from April 1997 through December 1997 using data for 1986 through 1994 on ISCR research files made available in March 1997. Procedures employed were as follows.

Procedures

1. The Hispanic origin, birthplace and race data on ISCR research files were first described. ISCR Hispanic origin data for 1986 to 1994 were compared with the Surveillance, Epidemiology, End Results Program (SEER) of the National Cancer Institute Hispanic cancer incidence data for 1988 to 1992 to determine differences in ancestral subgroup composition. In addition, ISCR cancer incidence files for 1986 to 1994 were matched to Illinois cancer mortality files for 1986 to 1996. Agreement between incidence and death certificate data for direct ascertainment of Hispanic origin was assessed to determine the quality of ISCR data relative to death certificate data.
2. Last and maiden names with identification and sequence numbers for cases each year were pulled from the Rocky Mountain Cancer Data System and then matched to identification and sequence numbers on the ISCR research files.

Then, a list of the surnames and schematic codes from the 1990 census Spanish surname study was matched electronically to the last and maiden names on the ISCR research files using the add variables procedure of SPSS for Windows (version 7.0) allowing surname classification of all ISCR cases according to the 1990 census Spanish surname study scheme. Those names not appearing on

the census study list were coded as 6000 (no match). Because the ISCR Rocky Mountain database is constantly updated and corrected, a small number of names was not available due to changes in case status and were missing for 1986 to 1993. These were coded as 9000 (unknown). Only 1994 was 100 percent complete for last names because names were pulled in close time proximity to the creation of the research file for this year. About 30 percent of females had maiden names available on the ISCR database. More than 90 percent would be expected given that cancer incidence predominates in older age groups. All missing maiden names were coded as 9000 (unknown). Thus, all last names for males and females and maiden names for females on the ISCR database were classified as one of 30 categories (28 U.S. Bureau of the Census study categories, no match category 6000 or missing category 9000).

3. Next, the most commonly heavily Hispanic (0125) and rarely Hispanic (5500) last names on ISCR for all cases, males and females were identified and compared with those from the 1990 census Spanish surname study for similarities and differences. ISCR maiden names in these categories also were evaluated.

4. Then, ISCR last names were compared with those from the census Spanish surname study as classified by the 28 categories, again to examine similarities and differences. Selected summary statistics including number of names in category, occurrences and percent Hispanic identification for all cases, males and females were the basis for comparisons. Although limited by incompleteness, the distribution of maiden names available on the ISCR database was examined similarly.

5. Next the annual distributions of the last names and maiden names by all 30 categories over the ISCR database from 1986 to 1994 were examined for number of occurrences, percent Hispanic, percent names per year and percent residing in Illinois Hispanic counties. In Illinois, most Hispanic persons reside in the following seven counties: Cook, DuPage, Kane, Lake, Rock Island, Will and Winnebago.

6. At this point in the study, the degree of similarity of the ISCR surnames to the census study surnames and their relationships to Hispanic status was determined. Depending on the findings, one of two approaches could have been taken with respect to the application of Spanish surnames to the identification of Hispanic cancer cases in Illinois. If the census study Spanish surname relationships with Hispanic status were similar to those observed for Illinois, the identification procedures would be done using the census Spanish surname schematic codes. If the comparison of the census study findings with Illinois were dissimilar, a Spanish surname schematic coding system tailored to Illinois would have been developed using surnames on the ISCR database for 1986 to 1994 and Hispanic origin as identified by cancer registrars using the census Spanish surname study methodology.

7. After it was determined that the best approach was to use the census Spanish surname study scheme rather than developing a separate scheme for Illinois, ISCR last and maiden names classified by the 30 categories were collapsed into seven broad categories including heavily Hispanic, generally Hispanic, moderately Hispanic, occasionally Hispanic, rarely Hispanic, no match and missing as follows.

0100 Heavily Hispanic: Categories 0101, 0102, 0105, 0110, 0115, and 0125
0200 Generally Hispanic: Categories 0201, 0202, 0205, 0210, 0215, and 0225
0300 Moderately Hispanic: Categories 0301, 0302, 0305, 0310, 0315, and 0325
0400 Occasionally Hispanic: Categories 0405, 0410, 0415, and 0425
5000 Rarely Hispanic: Categories 5001, 5005, 5020, 5025, 5100, and 5500
6000 No Match: Category 6000
9000 Missing: Category 9000

Percent Hispanic, percent names per year and percent residing in Hispanic counties were again examined by collapsed category for ISCR last and maiden names.

8. Hispanic surname variables were then created. First, cases born in places where Spanish names are prevalent but ethnic status is generally not Hispanic were excluded from Hispanic identification by name. Also, cases whose race was documented as American Indian, Filipino, or Hawaiian were excluded from Hispanic identification using name criteria. Hispanic identification using name was then implemented to produce four Hispanic identification variables that reflected the strength of association of name with Hispanic status for each case on the ISCR database (shown in Table 3).

In summary, males with last names in the 5000, 6000 and 9000 categories were not considered for Hispanic identification using Spanish surname classification. Available maiden names for females took priority over last names for Hispanic identification. Females with maiden names in the 5000 and 6000 categories were not considered for Hispanic identification regardless of the Spanish status of their last name. If maiden name was missing (9000) for females, last name was used for Hispanic identification. Females without maiden name information and last names in the 5000, 6000 and 9000 categories were excluded from Hispanic identification. Females with maiden names in the 0200, 0300 and 0400 categories were considered for Hispanic identification based on maiden or last name status as described in Table 3.

9. Then, a birthplace variable was created for all cases on the ISCR database for Hispanic identification purposes. Cases identified as Hispanic were those born in non-U.S. birthplaces associated with high probability of Hispanic ethnicity listed in Table 2. (Hispanic birthplace in this document refers to only non-U.S. birthplaces and does not include the 11 states where the Hispanic population is concentrated in the U.S.)

An adjustment that eliminated those cases classified as Hispanic by birthplace based on last and maiden name status was applied as follows. Males with last names in the 5000, 6000 or 9000 categories were eliminated from Hispanic identification based solely on birthplace. For females maiden name took priority over last name unless the maiden name was 9000 (unknown). That is, females born in non-U.S. Hispanic birthplaces with maiden names in the 5000 and 6000 categories were excluded from Hispanic identification based solely on birthplace. Females whose maiden names were 9000 (unknown) were classified based on last name using the same criteria that was applied to males. The adjustment eliminated non-Hispanic individuals who had settled in Spanish countries. (Many generations of English, German, Dutch and French families have resided in areas of South America.)

10. Hispanic identification was then implemented using available data on Hispanic origin, surname and birthplace. Fourteen Hispanic variables using combinations of available data were created and are shown in Table 4. The numbers and percent Hispanic cases identified over the ISCR database from 1986 to 1994 using each variable were calculated and compared to determine the effects of Hispanic status information separately and in combination.

11. The four combined Hispanic non-U.S. birthplace/surname variables (indirect Hispanic identifiers) were then compared relative to the Hispanic origin only variable (direct Hispanic identifier) over the 1986 to 1994 time frame as follows. The sensitivity, positive predictive value (PPV), specificity and negative predictive value (NPV) were calculated for each of the four Hispanic non-U.S. birthplace/surname combinations relative to ISCR Hispanic origin data for all cases in order to determine how relaxing the surname criteria through the graded inclusion of the 0100, 0200, 0300 and 0400 surname categories affected the indirect Hispanic identification relationship to Hispanic origin as defined by cancer registrars.

12. Three Hispanic identifier variables were selected to calculate and to evaluate cancer incidence rates for Illinois based on findings from evaluations of Hispanic non-U.S. birthplace, surnames and from comparisons of indirect to direct Hispanic identification on ISCR. These included a direct identifier, an indirect identifier, and a combined direct and indirect identifier.

Denominators used for cancer incidence rate calculations included the 1990 Modified Age, Race, and Sex (MARS) population counts for white Hispanic, non-white Hispanic, Hispanic (all races), and non-Hispanic (all races) males and females as well as U.S. Bureau of the Census population counts and intercensal estimates for years 1986 to 1994 for Hispanics and non-Hispanics in Illinois. Cancer incidence rates were calculated for selected sites using the Health Information Retrieval System (HIRS) software developed by the U.S. Centers for Disease Control and Prevention and were age-adjusted using the 1970 U.S. standard.

Cancer incidence rates were calculated for sites that were reported recently in the monograph, *Racial/Ethnic Patterns of Cancer in the United States 1988-1992*.⁸ These included all sites combined, oral cavity excluding nasopharynx, nasopharynx, esophagus, stomach, colon and rectum, colon, rectum, liver and intrahepatic bile duct, pancreas, larynx, lung and bronchus, melanomas, breast (invasive), urinary bladder, kidney and renal pelvis, brain and other nervous system, thyroid, Hodgkin's disease, non-Hodgkin's lymphoma, multiple myeloma and leukemias for males and females. Prostate and testis cancer incidences rates also were calculated for males and breast (*in situ*), cervix (invasive), corpus uteri and ovary for females. In addition, gallbladder cancer incidence rates for white Hispanics and white non-Hispanics were calculated, even though SEER rates were not reported, because gallbladder-related diseases have been shown to be more common among Hispanics, particularly Mexican Americans.⁹

ISCR Hispanic cancer incidence rates were calculated using incidence for three five-year periods including 1988 to 1992, 1989 to 1993, and 1990 to 1994 and compared with SEER white Hispanic and white non-Hispanic cancer incidence rates for 1988 to 1992. ISCR Hispanic (all races) cancer

incidence rates were also calculated for the three five-year periods (1988 to 1992, 1989 to 1993, and 1990 to 1994) and compared with reported Hispanic (all races) cancer incidence rates for SEER 1988 to 1992.

The purpose of calculating ISCR rates for three five-year periods was to examine comparability of ISCR to SEER while taking into account differences in the quality of Illinois cancer incidence data over time because, first, Hispanic identification in ISCR was implemented in 1989 and, second, completeness of the ISCR database has increased from inception of the registry in 1985 to the present. For these analyses, expected rates also were computed for the ISCR white Hispanic population based on the SEER white non-Hispanic, SEER white Hispanic, and the observed ISCR white non-Hispanic rates. Comparison of observed rates with expected rates for Hispanics allowed the examination of completeness effects of registry data on the Hispanic rates. Expected ISCR Hispanic rates were not calculated for Hispanics (all races) because cancer incidence rates for non-Hispanics (all races) were not reported by SEER and would not differ greatly from the white Hispanic/white non-Hispanic comparisons.

Means and standard deviations were computed for sites reported by SEER excluding all sites combined and combined colon and rectum. Pearson correlations were also made between the observed ISCR rates and the reported SEER rates for the each respective race-Hispanic-sex group as well as observed to expected ISCR rates for white Hispanics.

ISCR cancer incidence rates also were calculated for years 1986 to 1994 using the three Hispanic identification variables for the selected sites. The SEER Hispanic (all races) cancer incidence rates for 1988 to 1992 are presented with the ISCR Hispanic (all races) cancer incidence rates for 1986 to 1994.

13. Based on the results of this study, the methodology by which Hispanic cancer cases are defined in ISCR and used for all future division publications and studies was determined.

Results and Discussion

Tables 5 through 31, Appendices A and B, present results of the Hispanic identification study for ISCR, 1986 to 1994.

ISCR Hispanic Origin Data Element

Table 5 shows the annual distribution by Hispanic origin category for cancer cases on ISCR. It can be seen that the introduction of a separate Hispanic identification element in 1989 resulted in a dramatic increase in the identification of Hispanic cases by cancer registrars as well as in ancestral subgroup classifications. Further examination of categories in Table 5 shows some "Spanish surname only" reporting as well as a sizeable number of cases reported in the "unknown" category.

When the cases classified as "Spanish surname only" were examined after matching names with the census Spanish surname categories all cases, with the exception of three cases in the heavily Hispanic (0125) category, had names in the 5000 (rarely) or 6000 (no match) categories. Subsequent investigation by registry personnel found this to be the result of a Rocky Mountain Registry Data System conversion error when these cases were reported to ISCR by the Indiana Cancer Registry. Quality control procedures were then implemented to correct the problem in the registry database.

An examination of the surnames for cases classified as "unknown" revealed only 138 last names and 11 maiden names in the collapsed Hispanic categories 0100, 0200, 0300, or 0400 among the 3,022 "unknown" cases. Apparently, the criteria for cancer case classification as unknown Hispanic origin need clarity for cancer registrars. However, such classification would not influence ISCR analyses because "unknowns" become non-Hispanic for analytic purposes.

Mexican is the largest Hispanic ancestral subgroup followed by Puerto Rican (Table 6). Considerable improvement in the ability to discriminate subgroups within the ISCR Hispanic origin element (Table 6) does not seem evident with further refinement in 1993. A substantial proportion (range 36.2 to 45.5 percent) of cases were still classified as Hispanic, no other specification (NOS) in 1992, 1993, and 1994 when the revised incidence report form would have been used routinely.

ISCR Hispanic Ancestral Subgroup Distribution Compared with SEER 1988 to 1992

The Hispanic ancestral subgroup composition for SEER cancer incidence data 1988 to 1992 was obtained for comparison with ISCR data 1989 to 1994 for ancestral subgroup differences as shown in the following table.

Hispanic Category	ISCR 1989 to 1994		SEER ¹⁰ 1988 to 1992	
	count	%	count	%
<u>Mexican</u>	<u>2,021</u>	<u>0.67</u>	<u>3,362</u>	<u>0.48</u>

Puerto Rican	758	0.25	1,466	0.21
Cuban	240	0.08	247	0.04
South/Central American	323	0.11	1,282	0.18
Other Spanish (incl European)	141	0.05	4,465	0.63
Hispanic, NOS	2,319	0.77	13,912	1.98
Spanish Surname only	134	0.04	1	0.00
Unknown	3,000	0.99	18,071	2.57
Non-Hispanic	293,191	97.04	661,138	93.92
Total	302,127	100.00	703,944	100.00
				0

Hispanic Ancestral Subgroup	ISCR		SEER ¹⁰	
	1989 to 1994		1988 to 1992	
	count	%	count	%
Mexican	2,021	34.83	3,362	13.59
Puerto Rican	758	13.06	1,466	5.93
Cuban	240	4.14	247	1.00
South/Central American	323	5.57	1,282	5.18
Other Spanish (incl European)	141	2.43	4,465	18.05
Hispanic, NOS	2,319	39.97	13,912	56.25
All Subgroups	5,802	100.00	24,734	100.00

When all Hispanic categories are compared, SEER has a larger proportion of Hispanic cancer cases (approximately 6 percent) compared with ISCR (approximately 3 percent). The distribution by Hispanic ancestral subgroup shows considerably more Hispanic, NOS for SEER than for ISCR. Although it is difficult to compare the two cancer incidence populations because of their Hispanic, NOS differences, SEER probably has a greater proportion of Mexican cases than Illinois based upon the following excerpt from the SEER monograph, *Racial/Ethnic Patterns of Cancer in the United States 1988-1992*, which describes the Hispanic population of SEER.

Most of the Hispanic population in SEER lives in Los Angeles (60%), followed by New Mexico (10%), San Francisco and San Jose/Monterey (9%), and Connecticut (4%). About 84% of the Hispanic population in San Jose/Monterey identified themselves as Mexican-American in the 1990 census. Mexican-Americans account for 76% of the Hispanic population in Los Angeles, 58% in San Francisco/Oakland and 57% in New Mexico. San Francisco/Oakland also has a sizable percentage of Puerto Rican Hispanics (4%). Over two-thirds of the Hispanic population in Connecticut is Puerto Rican, with smaller percentages of Mexican-Americans (4%) and Cubans (3%).⁸

From this excerpt, it might be assumed that a greater proportion of those SEER Hispanic, NOS cases shown in the above table are, in fact, Mexican. Thus, the ISCR Hispanic cancer incidence population is probably composed of less Mexican and other Spanish, including European, and more Puerto Rican and Cuban Hispanic ancestral subgroups than the SEER Hispanic cancer incidence population.

ISCR Hispanic Identification for Agreement Compared with Illinois Death Certificates

A final examination of the quality of the Hispanic origin variable on the ISCR database involved comparison of ISCR cases matched to Illinois master death files. The 437,099 cancer incidence cases for 1986 to 1994 were matched to 267,499 cancer deaths occurring from 1986 to 1996 on last name and social security number using the add variables procedure of SPSS for Windows (version 7.0). A file with 142,309 matched cases was produced. Analyses for agreement between cancer incidence records and death certificate records were conducted for Hispanic origin using matched records for the 93,984 matched records with cancer cases diagnosed from 1989 to 1994. The results are summarized as follows.

Analysis for agreement between the Hispanic origin element on the ISCR incidence report form with Hispanic origin on the death certificate showed the Kappa statistic to be 0.791 ($p < 0.001$) indicating 79.1 percent agreement beyond chance. Kappa statistics for the three major Hispanic ancestral subgroups, Mexican, Puerto Rican, and Cuban, were 0.565, 0.708 and 0.690, respectively. Agreement between the two Hispanic origin assessments may be considered excellent for all Hispanics and fair to good for each of the ancestral subgroups.¹¹

Birthplace

The reported birthplaces for cases identified as Hispanic by cancer registrars are shown in Table 7. Birthplace data are quite incomplete since the greatest proportion of cases (38.2 percent) have birthplace documented as unknown. However, more than 20 percent of Hispanic cases were reported to be born in Mexico. Illinois, Puerto Rico, Texas and Cuba were birthplaces reported in the next greatest proportions for ISCR Hispanic cancer cases.

Table 8 displays data for ISCR cases born in Hispanic birthplaces outside of the U.S. As would be expected, Mexico, Puerto Rico and Cuba had the largest proportions of Illinois cancer cases born in Hispanic birthplaces. The proportion of cases identified as Hispanic by cancer registrars within non-U.S. Hispanic birthplaces greatly increased when the Hispanic origin element was included on ISCR incidence report forms in 1989. From 1990 through 1994, the positive predictive value of all Hispanic birthplaces combined for Hispanic status identified by the ISCR Hispanic origin element exceeded 90 percent.

Table 9 shows cross-tabulations of Hispanic subgroups by race for ISCR cases. Approximately 97 percent of Hispanic cancer cases were classified as white race. Approximately 1 percent was reported as black race and the remainder was distributed among other racial groups. The largest proportions of Hispanic cancer cases reported to be black race were Puerto Rican and Cuban, which would be expected given that historically African slaves were brought to these islands.

Comparison of ISCR Surnames with Census Spanish Surname Study

Table 10 shows a comparison of last names for all cases on the ISCR database with those identified in the 1990 census Spanish surname study by rank for the 12 most common heavily Hispanic surnames (0125) and the 12 most common rarely Hispanic surnames (5500). As shown, the ISCR names are remarkably similar to those observed for the census study. With a few exceptions, only slight differences in rank were apparent. The name Gonzales ranked 12th for the census sample and 30th on the ISCR database among heavily Hispanic names. Among the rarely Hispanic names, Jackson ranked 12th on the ISCR database and 20th on the census data set. This difference most likely reflects the fact that cancer incidence rates are high among black males and the name Jackson is prevalent among the black population. (Comparisons for males, females and the most common maiden names for ISCR are shown in Appendix A Tables 1, 2 and 3.)

It is apparent that, at least from a frequency perspective for the most heavily and rarely Hispanic categories, the surnames on the ISCR database for years 1986 to 1994 are very similar to those observed in the census national sample. Of the 639 most frequently observed heavily Hispanic surnames (0125) in the 1990 census Spanish surname study sample, 574 appeared as last names on the ISCR database over 1986 to 1994.

Table 11 presents a comparison of selected summary statistics (number of surnames, occurrences and percent Hispanic) for ISCR last names with the census bureau Spanish surname study for all 28 Spanish surname categories. Percent Hispanic for ISCR cases reflects Hispanic origin data for years 1990 through 1994 when the Hispanic origin element was in full implementation, whereas the census Spanish surname study uses self-identified responses to the Hispanic origin question on the census form.

Fewer names were identified for most categories on ISCR than the census bureau Spanish surname study, especially for the lower frequency census scheme categories (i.e., 0101, 5001) due to the fact that the number of occurrences on the ISCR database are considerably less than those in the census study sample. Predictions of surname categories for Hispanic status showed similar patterns for both ISCR and the census study sample. That is, greater proportions of Hispanics are identified as the surname category becomes more Hispanic. However, the magnitudes of the surname category Hispanic relationships were less for ISCR than for the census study. (e.g., persons with last names in the heavily Hispanic category 0125 showed 94.2 percent of the census bureau study sample to be Hispanic but only 73.2 percent ISCR cases with last names in the heavily Hispanic category 0125 were identified by cancer registrars as Hispanic.) Overall, the precision of the surname for predicting Hispanic status was better for males than for females on ISCR (see Table 11 heavily Hispanic categories) which would be expected given that the census bureau Spanish surname study findings are more generalizable to males and females who have never married.

The most likely explanation for differences in the surname category Hispanic relationships between the census sample and ISCR is that ISCR data are dependent on second-hand information. That is, cancer registrars are limited to medical records, which may not have the necessary information to assess Hispanic status accurately. The census bureau Hispanic origin question depends upon self-identification, which is most likely more precise than ISCR incidence report data.

Summary statistics for ISCR maiden names presented in Appendix A Table 4 show considerably fewer names and less occurrences than ISCR last names. In general, the expected surname category relationships to Hispanic status also are evident for the ISCR maiden names.

For Illinois, it appears that the census bureau Spanish surname classification system offers a more accurate means to indirect Hispanic identification than a scheme developed using only ISCR data. It is very unlikely that the true strength of the surname categorical relationships with Hispanic status would differ that greatly for the Illinois population compared with the census bureau study sample given the similarities in occurrence frequency and surname Hispanic status association patterns. Therefore, the study proceeded using ISCR surnames classified by the census bureau Spanish surname study classification scheme to evaluate indirect Hispanic identification using surname on the ISCR database.

ISCR Surnames by Collapsed Census Spanish Surname Study Categories

Tables 12 and 13 present selected summary statistics for ISCR last names and maiden names by collapsed census bureau Spanish surname study categories, respectively. More than 95 percent of last names on the ISCR database for 1986 to 1994 were among the rarely Hispanic or no match categories, about 0.5 percent of the names were missing, and the remaining 4.5 percent were in the heavily, generally, moderately, or occasionally Hispanic categories (Table 12). For females, Table 13 shows more than 70 percent missing maiden names, about 30 percent in the rarely Hispanic or no match categories, and about 0.6 percent in the heavily, generally, moderately, or occasionally Hispanic categories. (Selected summary statistics for all 28 categories for ISCR last and maiden names are presented in Appendix A Tables 5 and 6, respectively.)

For 1989 to 1994, when formal Hispanic identification was operational in ISCR, greater proportions of Hispanic cases are observed as the surname category becomes more Hispanic, progressing from occasionally Hispanic to heavily Hispanic, as well as the proportion of cases residing in Hispanic counties for both last and maiden names. In Table 12, it is again apparent that, on the ISCR database, name is a better predictor of Hispanic status for males than for females. Tables 12 and 13 also include the percentage of cases identified as Hispanic by surname classification with and without adjustment for birthplace and race. The proportion of cases excluded from Hispanic identification using surname based on birthplace and/or race ranged from 2.6 percent to 5.1 percent. For both last and maiden names, adjustment for birthplace and race enhanced the strength of the census Spanish surname categorical associations with Hispanic status in the expected direction supporting the use of adjustment to improve precision in the application of surnames for indirect Hispanic identification when birthplace and race data are available.

Surnames for Cancer Cases Born in Hispanic Non-U.S. Birthplaces

Table 14 shows the last names of ISCR cancer cases born in Hispanic birthplaces by the collapsed census Spanish surname study categories. As shown, approximately 80 percent of these cases have last names in the heavily Hispanic (0100) category. It is important to note that about 4 percent have names in the rarely Hispanic (5000) category and about 10 percent have names in the no match (6000) category. As presented in Tables 12 and 13, the proportion Hispanic for these categories was very low and, therefore, cases with surnames in these categories should be excluded when cases born in Hispanic birthplaces are indirectly identified as Hispanic.

American Indian, Filipino and Hawaiian Cancer Case Last Names

A total of 348 of 437,399 cases on the ISCR database from 1986 to 1994, or 0.08 percent, were reported to be among the three race groups where Spanish names are prevalent but Hispanic ethnicity is not common (Table 15). This combined race group included 306 (87.9 percent) Filipinos, 30 (8.6 percent) American Indians and 12 (3.4 percent) Hawaiians. Only Filipinos and American Indians had last names in the heavily, generally, moderately, or occasionally Hispanic surname categories. Among Filipinos 55 percent and among American Indians 20 percent of the last names were among the heavily, generally, moderately, or occasionally Hispanic surname categories, thus supporting the exclusion of cases based on race when indirect Hispanic identification procedures are implemented.

Effects of Surname Adjustment on Hispanic Non-U.S. Birthplace As Hispanic Identifier

An Hispanic birthplace variable was created for each case on the ISCR database for 1986 to 1994. When cases born in Hispanic non-U.S. birthplaces were excluded according to the Spanish surname criteria described previously in this report (see “Methods” p.3), the positive predictive value of Hispanic non-U.S. birthplace for Hispanic status increased from 92.7 percent to 96.8 percent for years 1990 through 1994 as shown below.

Hispanic Birthplace	Proportion Identified as Hispanic for Unadjusted and Adjusted Hispanic Non-U.S. Birthplace Illinois State Cancer Registry, 1986 to 1994									
	1986-94	1986	1987	1988	1989	1990	1991	1992	1993	1994
	No Formal Identification				ISCR Hispanic Identification					
Unadjusted	64.6	3.0	3.4	13.1	78.4	92.9	94.2	92.4	91.9	92.0
Adjusted	67.5	3.4	3.8	11.9	81.3	96.1	97.9	97.5	97.1	95.5

Such an adjustment serves to improve the precision of Hispanic identification using birthplace especially when only indirect evidence for Hispanic status is available and should be incorporated into algorithms for Hispanic identification.

Table 16 displays the comparison of the four combinations of Spanish surname/Hispanic non-U.S. birthplace relative to Hispanic origin as defined by cancer registrars. As the surname criteria becomes more relaxed to include cases with Spanish names in the generally, moderately, and occasionally

Hispanic categories, sensitivity increases, PPV decreases, specificity decreases and NPV increases. Although the decrease in specificity may appear slight, it should be kept in mind that far more cases are non-Hispanic than Hispanic on the database and what may appear as small, even a fraction of a percent, could translate into substantial false Hispanic identification of cancer cases with a drop in specificity. The use of indirect measures for cancer case ascertainment purposes should maximize correct identification. In the census bureau's working paper, this is referred to as minimizing the chance of making an "error of commission" or identifying a cancer case by indirect methods as Hispanic when the case is in reality not Hispanic.⁶ In epidemiologic terms, it is desirable to minimize "false positive" selections so that cancer incidence rates reflect the Hispanic population under evaluation.

Appendix A Tables 7 and 8 present the comparisons of the four combinations of Spanish surname/Hispanic non-U.S. birthplace relative to Hispanic origin for males and females, respectively. As would be expected, the relationships are stronger for males than for females because the generalizability of the census bureau Spanish surname approach is greatest for males and females who never married. In addition, the lack of completeness of maiden names on the ISCR database reduces the accuracy of Hispanic identification using surname.

Hispanic Identification

Table 17 shows the results of Hispanic identification using 14 combinations of Hispanic origin, Hispanic non-U.S. birthplace, and surname. It is evident that with additional data the number and proportion of cancer cases identified as Hispanic increases. The goal in evaluating these identifiers was to determine which maximize the selection of true Hispanic cases and minimize incorrect case selection.

Highlighted are the three identifiers selected for the study of Hispanic cancer incidence rates. The selected Hispanic identifiers were (1) ISCR Hispanic origin only as determined by cancer registrars, the direct identifier; (2) heavily Hispanic surname and/or Hispanic non-U.S. birthplace, the indirect identifier; and (3) ISCR Hispanic origin status and/or heavily Hispanic surname and/or Hispanic non-U.S. birthplace, the combined direct and indirect identifier.

ISCR Hispanic origin is a completely direct Hispanic identifier because these cases were determined to be Hispanic by cancer registrars through medical record reviews in accordance with data acquisition procedures established by ISCR. This means of identification has the distinct advantage of identifying Hispanic cancer cases when no evidence of Hispanic status is apparent through birthplace or surname.

Although the ISCR database has a means of directly identifying Hispanic cases, an indirect Hispanic identifier was selected to use in this study to determine the quality of cancer incidence rates produced through indirect identification. Ideally, an optimal indirect Hispanic identifier would be established with an algorithm that maximizes correct identification using birthplace, race, marital status and last or maiden names. With the exception of marital status, all of these variables are

available on the ISCR database and were used to create the identifiers used to produce data shown in Tables 16 and 17.

The Hispanic identifier that included cases born in Hispanic non-U.S. birthplace and/or cases with heavily Hispanic surname was selected for further evaluation in this study as the indirect identifier. The rationale for selection of this identifier was that, first, cases born in Hispanic birthplaces adjusted for surname have almost a 97 percent probability of being Hispanic. Second, the heavily Hispanic surnames were shown to be associated with the greatest proportions of Hispanic persons for both ISCR and the census Spanish surname study (Table 11). For the census study, the heavily Hispanic (0100) categories were over 90 percent Hispanic and for ISCR between 50 percent to 75 percent Hispanic with the exception of category 0101. Appendix B contains an alphabetical list of the 0100 last names appearing on the ISCR database from 1986 to 1994 with the specific census bureau Spanish surname category.

Although the census bureau study sample of generally Hispanic 0200 categories also contained high proportions of Hispanic persons (50 percent to 77.9 percent) percent Hispanic, the proportions Hispanic in the 0200 Spanish surname categories for ISCR were much lower (7.0 percent to 39.9 percent). In addition, although the national census sample and the ISCR population appear to be very similar based on evaluation of surnames, differences undoubtedly do exist and those difference most likely would be manifest in the generally, moderately and occasionally Hispanic surname categories. An example of this was apparent when all surnames in the heavily, generally, moderately, and occasionally Hispanic surname categories were examined for frequency of occurrence on the ISCR database. The surname Moran, category 0425, appeared among the top 12 when all heavily, generally, moderately and occasionally surnames on ISCR were considered. Moran is occasionally common among Hispanics nationally but is also prevalent among persons of Irish origin and Illinois has a large Irish population subgroup. If all of the Hispanic category surnames were considered, substantial false positive identification would have been made because Moran would have been used as an indirect identifier. For these reasons, it was decided that only heavily Hispanic surnames in the (0100) categories be used for indirect Hispanic identification on ISCR to reduce the chance of error.

The third Hispanic identifier is the strongest identifier among those presented in Table 17 because it combines both direct and indirect information in a manner that maximizes correct Hispanic identification based on the results of this evaluation. “Hispanic” is defined as (1) all cases identified by cancer registrars as Hispanic; (2) cases with heavily Hispanic surnames (last name for males, maiden name for females, last name for females if maiden name not available) not born in birthplaces associated with high Spanish surname prevalence and low probability of Hispanic ethnicity and not of American Indian, Filipino or Hawaiian race; and (3) cases born in birthplaces associated with high probability of Hispanic ethnicity and not having maiden name for females in the rarely Hispanic or no match categories and for males and females without maiden names whose last names appear in the rarely Hispanic, no match or missing categories. In this manner, all information on the ISCR database is used in a manner that maximizes accurate Hispanic identification and minimizes false positive identification.

Table 18 shows the proportion Hispanic cancer deaths identified using each of the 7 possible combinations of Hispanic origin, Hispanic non-U.S. birthplace and heavily Hispanic last or maiden name when Hispanic cancer cases are identified using the combined direct and indirect identifier. Between 15 percent to 20 percent of cases could only be identified directly by cancer registrars and slightly more than 20 percent were identified indirectly using Hispanic non-U.S. birthplace and/or heavily Hispanic surname for years 1990 to 1994 when the ISCR Hispanic origin data element was fully implemented. As shown, most identification (75 percent to 80 percent) was accounted for using Hispanic origin as documented by cancer registrars and cases with heavily Hispanic surnames. Hispanic non-U.S. birthplace alone identified the least Hispanic cases when all sources for identification are employed.

It should also be noted that indirect Hispanic identification does not compensate for the absence of formal Hispanic identification for years 1986 to 1988 on ISCR. However, it does appear to provide an additional approach to find Hispanic cases not identifiable by information available to cancer registrars.

It should also be kept in mind that even though an optimal identifier has been defined, some misclassification of cases still exists for the following reasons. First, errors occur during registry operations that result in documentation of Hispanic cases as non-Hispanic and vice versa. Second, the positive predictive values of Hispanic non-U.S. birthplaces and heavily Hispanic surnames for Hispanic status were shown to be high but were not 100 percent even after adjustments. Third, true Hispanic cases with surnames in the generally, moderately, occasionally and no match categories that are not identified by cancer registrars cannot be ascertained indirectly based on the definition established for indirect identification. However, the number of these cases is probably so small that assessment of cancer incidence among Hispanics in Illinois would most likely not be adversely affected.

Hispanic Cancer Incidence Rates

Tables 19 through 24 display white Hispanic and white non-Hispanic cancer incidence rates for ISCR using the three Hispanic identifiers compared with those available from SEER. The SEER Hispanic rates represent one of the few standards for comparison. However, when using SEER as a standard for comparison, the assumption must be made that the Illinois Hispanic population is similar to the SEER Hispanic population with respect to cancer incidence. It was shown previously on page 11 that some differences between ISCR and SEER were apparent for Hispanic ancestral subgroup composition and should be kept in mind when interpreting data. Although the known patterns of cancer for Hispanics tend to be shared by most of the ancestral subgroups, the fact that Illinois' Hispanic composition differs somewhat from SEER should be recognized when interpreting these data.

In each table, ISCR rates are calculated using cancer incidence for 1988 to 1992, 1989 to 1993 and 1990 to 1994. Expected rates for ISCR white Hispanics also have been calculated as described under "Methods" on page 9. Tables 19, 20 and 21 present comparisons for white Hispanic and white

non-Hispanic males identified using the direct identifier, the indirect identifier and the combined direct and indirect identifier, respectively. Tables 22, 23 and 24 present the same comparisons for white Hispanic and white Non-Hispanic females.

Tables 25, 26 and 27 present descriptive and analytic statistics for comparisons of cancer incidence rates calculated for ISCR with those reported by SEER 1988 to 1992 for white Hispanics and white non-Hispanics. Pearson correlation coefficients indicated that comparisons for the nine white Hispanic ISCR cancer incidence rates (three Hispanic identifiers x three ISCR incidence time periods) with the SEER 1988 to 1992 rates were highly correlated ($p < 0.01$) for both males and females.

However, the ISCR white Hispanic cancer incidence rates calculated using the combined direct and indirect Hispanic identifier for years 1990 to 1994 when the Hispanic origin data element was in full implementation and the registry database was more complete (Table 21 for males and Table 24 for females) consistently had the highest Pearson correlation coefficients when compared with SEER and the means of site-specific cancer incidence rates were closest to those for SEER. ISCR cancer incidence rates calculated with the direct Hispanic identifier were lowest when compared with the SEER rates even when ISCR data for years 1990 to 1994 were used to calculate rates (Tables 19 and 22). ISCR cancer incidence rates calculated using the indirect Hispanic identifier (Tables 20 and 23) were between those produced with the direct Hispanic identifier and those produced using the combined direct and indirect Hispanic identifier.

As shown in Table 27, comparisons of observed ISCR rates to expected ISCR rates (Pearson correlation coefficients range 0.992 to 0.998) were highly correlated ($p < 0.01$) for all three Hispanic identifiers. The observed means compared with expected means for rates calculated using the combined identifier showed the least difference in all three time periods for both males and females.

The patterns of cancer incidence among Hispanics compared with non-Hispanics in Illinois appear to be consistent with the available reports on Hispanic cancer incidence.^{8,12,13} Hispanic males and females had lower all sites combined cancer rates compared with their non-Hispanic counterparts in Illinois. Hispanic males in Illinois have lower cancer incidence rates of lung, prostate, colon, rectum, oral cavity, melanoma and non-Hodgkin's lymphoma than non-Hispanic males in Illinois. Hispanic females in Illinois have lower rates of breast, colon, rectum, corpus uteri, lung, ovary, oral cavity, non-Hodgkin's lymphoma and melanoma than their non-Hispanic counterparts. Conversely, Hispanic males have higher rates of stomach, liver and gallbladder compared with non-Hispanic males in Illinois. Higher rates of cancers of the cervix, stomach, liver and gallbladder occurred among Hispanic females than non-Hispanic females in Illinois. These patterns were apparent for cancer incidence rates calculated with each of the three Hispanic identifiers.

Tables 28 and 29 show five-year Hispanic (all races) cancer incidence for ISCR with SEER 1988 to 1992 for males and females, respectively. Cancer incidence rates also were calculated for the three Hispanic variables using ISCR data for 1988 to 1992, 1989 to 1993 and 1990 to 1994. Comparisons of ISCR Hispanic cancer incidence rates with SEER showed the same patterns that were described

for white Hispanics and are supported by descriptive and analytic statistics presented in Table 30. The ISCR combined direct and indirect Hispanic identifier for years 1990 to 1994 produced cancer incidence rates least different from those reported for SEER 1988 to 1992. SEER non-Hispanic (all races) cancer incidence rates were not available to calculate expected ISCR rates for these comparisons.

The most frequently occurring cancer sites for Hispanic males and females in Illinois were consistent with those observed for SEER and other Hispanic populations in the United States.^{8,12,13} Prostate, lung, colon, kidney, non-Hodgkin's lymphoma, stomach and rectum were the most common sites of cancer for Illinois' Hispanic males. For Illinois' Hispanic females, cancers of the breast, lung, cervix, colon and corpus uteri were most common.

Tables 31 and 32 show Hispanic cancer incidence rates across the ISCR database time frame from 1986 to 1994 using the three Hispanic identifiers for males and females, respectively. The SEER Hispanic cancer incidence rate for 1988 to 1992 also is presented with each site for comparison purposes.

The implementation of the Hispanic identifier in 1989 is most apparent for the direct identifier (ISCR Hispanic origin only) and also for the combined direct and indirect identifier when examining rates from 1986 to 1994, especially for all sites combined. These cancer incidence rates for 1986 to 1988 are substantially lower than those from 1989 to 1994 illustrating greater error in the earlier years or less identification before Hispanic origin was ascertained directly in 1989. Unequal amounts of error across a time frame interferes with the ability to examine trends. Therefore, if direct means alone or in combination with indirect sources are used to identify ISCR Hispanic cancer cases, cancer incidence rates should only be reported from 1989 and beyond.

In contrast, use of the indirect Hispanic identifier results in consistent error over the 1986 to 1994 time frame. However, true Hispanic cancer cases that were not born in Hispanic birthplaces and do not possess heavily Hispanic surnames are missed using indirect means and the cancer incidence rates produced would be less accurate than when a direct identifier is available for ascertainment of Hispanic cancer cases. If only indirect sources were available for Hispanic identification, cases with generally Hispanic surnames might be included as Hispanic to compensate for the loss of cases in the absence of direct methods.

Conclusions

1. Hispanic identification was effectively implemented in the Illinois State Cancer Registry in 1989 when a specific data element was included on the data acquisition incidence report form. Agreement between ISCR and Illinois death certificate Hispanic identification was nearly 80 percent.
2. Surnames and their relationships to Hispanic status categorically presented in the U.S. Bureau of the Census working paper, *Building a Spanish Surname List for the 1990's A New Approach to An Old Problem*, were very similar to those observed on the ISCR database for 1986 to 1994 and offer an indirect means for Hispanic identification in Illinois.
3. Hispanic non-U.S. birthplaces including Central America (Guatemala, Belize, Honduras, El Salvador, Nicaragua, Costa Rica and Panama), Cuba, Mexico, Puerto Rico, South America (Colombia, Venezuela, Ecuador, Peru, Bolivia, Chile, Argentina, Paraguay and Uruguay) and Spain (Canary Islands, Balearic Islands and Andorra) were demonstrated to be valid indirect identifiers of Hispanic status.
4. Direct Hispanic identification using the data acquisition element implemented in 1989 was shown to have the capability of finding about 75 percent to 80 percent of Hispanic cancer cases. The remaining 20 percent to 25 percent could be identified using Hispanic non-U.S. birthplace and heavily Hispanic surnames. About 15 percent of all Hispanic cancer cases could only be identified directly by cancer registrars.
5. Exclusion of cases based on race, birthplace and/or surname status from indirect identification to improve precision was shown to increase positive predictive values of Hispanic non-U.S. birthplaces and Hispanic surnames for Hispanic status. Therefore, these exclusions should be incorporated into algorithms for Hispanic identification on ISCR.
6. Indirect identification was shown to augment accurate Hispanic identification on ISCR and should be incorporated routinely into data analyses and evaluation procedures.
7. ISCR cancer incidence rates calculated using Hispanic cases identified with a combined direct and indirect Hispanic identifier for years 1990 to 1994 showed the least difference from SEER for Hispanics.
8. Based on the findings of this study, the Hispanic identifier that will be used routinely on ISCR is as follows. "Hispanic" is defined as (1) all cases identified by cancer registrars as Hispanic; (2) cases with heavily Hispanic surnames (last name for males, maiden name for females, last name for females if maiden name not available) not born in birthplaces associated with high Spanish surname prevalence and low probability of Hispanic ethnicity and not of American Indian, Filipino or Hawaiian race; and (3) cases born in birthplaces associated with high probability of Hispanic ethnicity and not having maiden name for females in the rarely Hispanic or no match census Spanish surname study categories

and for males and females without maiden names whose last names appear in the rarely Hispanic, no match or missing categories.

9. ISCR cancer incidence will be reported starting from 1989 when the Hispanic origin data element was established on the registry cancer incidence report form.

Limitations

1. All results from this study are only estimates because it is not possible to know for certain the exact number of Hispanic cancer cases in Illinois. Undoubtedly, a small proportion of Hispanic cancer cases are missed even though procedures of direct identification by cancer registrars and indirect identification through birthplace, race and surname are employed routinely. Despite efforts to avoid false positive identifications, some cases will still be identified as Hispanic that are indeed non-Hispanic since the indirect identification criteria are not 100 percent certain and misclassifications also may occur during direct identification due to errors in registry operations. Even though it is not possible to estimate how many cases are missed or misclassified, we feel that the actual numbers are small using the combined direct and indirect Hispanic identifier developed in this study.
2. Identification using the algorithm, which incorporates both direct and indirect methods, cannot specify Hispanic ancestral subgroup. Spanish surnames are not specific to ancestral subgroups.
3. The surname approach from the U.S. Bureau of the Census is applicable for the 1990s. As Hispanics acculturate, changes in surname Hispanic status relationships will occur nationally and in Illinois. These changes may render use of Spanish surname lists obsolete and, therefore, the effectiveness of algorithms for Hispanic identification should be monitored and changed accordingly.
4. It is fortunate that the census Spanish surname study approach could be applied in its entirety to the cancer incidence population of Illinois. The ability to use this application is attributable to the similarity of the Illinois Hispanic population to the nation's Hispanic population. States sharing this representativeness could probably use the approach as has Illinois. However, other states, such as New Mexico, whose Hispanic surnames are unique would have more difficulty in its application.⁵ This factor should be considered when using Spanish surname approaches to indirect Hispanic identification.

References

1. Census Bureau releases refined estimates from post-enumeration survey of 1990 census coverage. *U.S. Department of Commerce News*. June 13, 1991. U.S. Dept of Commerce Publication CB #91-221.
2. Council on Scientific Affairs, American Medical Association. Hispanic health in the United States. *J.A.M.A.* 1991;265:248-252.
3. United States Bureau of the Census, *Estimates of the Population of Counties by Age, Sex and Race/Hispanic Origin: 1990 to 1996*. Population Division, 1997.
4. United State Bureau of the Census, *General Characteristics of Persons by Spanish Origin and Race, 1980*. Illinois, Population Division.
5. North American Association of Central Cancer Registries. Uniform Data Standards Committee. *Subcommittee on Methodologic Issues of Measuring Cancer among Hispanics. Final Report of Atlanta Symposium*, 1996.
6. Seiffert JE (ed). *Standards for Cancer Registries, Vol II, Data Standards and Data Dictionary*. Sacramento, Calif.: North American Association of Central Cancer Registries, March 1997.
7. Word DL and Perkins RC. *Building a Spanish Surname List for the 1990's--A New Approach to An Old Problem*. Population Division Working Paper No. 13. February 1996.
8. Miller BA, Kolonel LN, Bernstein L, Young, Jr. JL, Swanson GM, West D, Key CR, Liff JM, Glover CS, Alexander GA, et al. (eds). *Racial/Ethnic Patterns of Cancer in the United States 1988-1992*, National Cancer Institute. NIH Pub. No. 96-4104. Bethesda, MD, 1996.
9. Wagener DK and McDonald M. Increased gallbladder-related mortality among Hispanics: does education Play a role? *Ethn Health* 1996;197-205.
10. SEER 1973-1994 Public-Use CD-ROM. The National Cancer Institute. (Issued May 1997).
11. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977a;33:159-174.
12. Trapido EJ, Valdez RB, Obeso JL, Strickman-Stein N, Rotger A, Perez-Stable EJ. Epidemiology of cancer among Hispanics in the United States. *Monogr Natl Cancer Inst* 18:17-28, 1995.

13. Gutierrez-Ramirez A, Valdez RB, Carter-Pokras O. Chapter 8. Cancer. In: Molina CW Aguirre-Molina M. Eds. *Latino Health in the U.S.: A Growing Challenge*. American Public Health Association, 1994.