Recalibrating the NAACCR Hispanic Identification Algorithm (NHIA)

Xiuling Zhang, Francis P. Boscoe, New York State Cancer Registry

Introduction

For over a decade, NAACCR member registries have been using the NAACCR Hispanic Identification Algorithm (NHIA) to enhance the completeness and accuracy of Hispanic ethnicity coding. The algorithm uses country of birth, surname, and county of residence to modify the reported Hispanic code.

Prior to the existence of NHIA, the New York State Cancer Registry (NYSCR) developed its own Hispanic identification algorithm (NY-NHIA), which it has continued to use through the present, and which differs from NHIA in several minor respects. New York's conversion to SEER*DMS in the spring of 2016 prompted a re-evaluation of these differences to see if the two versions could be brought into alignment.

During this process, we became aware that all NAACCR member registries, including New York, were still using a surname list developed in the 1990s that was based on very a small sample size. We expanded our project to assess the impact of substituting a recently published population-based* surname list.

This poster reports on the differences in Hispanic cancer rates between the NHIA and NY-NHIA and between the sample-based and population-based surname lists, for New York State, several other state registries, and the nation as a whole.

The primary differences between NHIA and NY-NHIA are:

1. With respect to birthplace, NHIA prioritizes the reported Hispanic code, so that if a patient was reported as Puerto Rican, then the patient is counted as Puerto Rican, even if he was born in Cuba. NY-NHIA prioritizes the reported birthplace, so that the above patient would be counted as Cuban. This difference has no impact on Hispanic rates, only on the distribution of Hispanic subgroups.

2. NHIA checks the surname list for patients reported as non-Hispanic, so that Jesús Figueroa would be counted as Hispanic, even if reported as non-Hispanic. NY-NHIA does not check the surname list for patients reported as non-Hispanic, so that Jesús Figueroa would be counted as non-Hispanic, if so reported. This difference results in lower Hispanic rates when using NY-NHIA.

There are also some additional minor differences, too nuanced to list here.

Methods

- ❑ All invasive malignant cancer cases diagnosed between 2009 and 2013 were obtained from the NYSCR (n=505,601).
- ❑ The NHIA and NY-NHIA algorithms were applied to all the NY cases, with all surnames changed to "SMITH" exclude the surname component.
- ❑ The NHIA and NY-NHIA algorithms were then applied to all the NY cases using both the 1990 and 2010 Hispanic surname lists.
- ❑ The exercise was repeated with data from 44 other registries (using the CINA database) to evaluate the birthplace component and 3 volunteer registries to evaluate the surname component.
- ❑ All resulting differences were identified and tabulated.

Results & Discussion

Table 1. Comparison of NHIA and NY-NHIA using New York data

NY-NHIA	NHIA								
	0	1	2	3	4	5	6	8	Total
0	460,787	16	12	3	117	154	289	4	461,382
1	0	1,530	3	1	20	24	0	0	1,578
2	0	7	9,457	0	37	242	0	10	9,753
3	0	4	3	1,043	7	35	0	0	1,092
4	0	16	32	4	7,156	1,001	0	9	8,218
5	0	0	2	0	3	4,211	0	2	4,218
6	0	0	0	0	0	0	13,776	0	13,776
8	0	3	39	2	223	900	0	4,417	5,584
Total	460,787	1,576	9,548	1,053	7,563	6,567	14,065	4,442	505,601

As shown in Table 1, the algorithms disagreed 0.64% of the time (3,224/505601). The two most common disagreements, representing about half of the differences, involved cases coded as "other specified Spanish/Hispanic origin" (code 5) by NHIA and either South or Central American (code 4) or Dominican Republic (code 8) by NY-NHIA (highlighted in yellow). This suggests tumor registrars in New York are occasionally failing to recognize that more specific codes are available than 5.

NY-NHIA resulted in 44,219 cases coded as Hispanic; NHIA resulted in 44,814 (differences highlighted in red). Crude Hispanic cancer rates using New York's algorithm are reduced by the ratio of these two numbers (1.3%).

Table 2. Comparison of NHIA and NY-NHIA using CINA data for 44 registries excluding New York

NY-NHIA	NHIA								
	0	1	2	3	4	5	6	8	Total
0	5741213	91	46	11	261	407	883	12	5742924
1	0	82674	16	3	21	171	0	4	82889
2	0	34	17828	7	38	75	0	5	17987
3	0	28	16	16313	14	34	0	2	16407
4	0	108	29	21	29449	468	0	5	30080
5	0	6	1	2	6	8014	0	0	8029
6	0	0	0	0	0	0	214817	0	214817
8	0	2	8	3	136	77	0	4647	4873
Total	5741213	82943	17944	16360	29925	9246	215700	4675	6118006

As shown in Table 2, the algorithms disagreed just 0.05% of the time (3,051/6,118,006).

The use of code 5 for patients born in South/Central America is the most common discrepancy (yellow highlight), but the total number is less than half of what is seen in New York alone.

NY-NHIA resulted in 375,082 cases coded as Hispanic; NHIA resulted in 376,793 (differences highlighted in red). Crude Hispanic cancer rates using New York's algorithm would be reduced by the ratio of these two numbers (0.5%).

Table 3. Comparison of 1990 and 2010 Hispanic surname lists

1990 Surname List	2010 Surname List				
	Heavily Hispanic	Rarely Hispanic	On list, but neither heavily nor rarely	Not on list	Total
Heavily Hispanic	4,393	90	949	6,783	12,215
Rarely Hispanic	45	4,782	1,871	1,515	8,213
On list, but neither heavily nor rarely	440	206	1,308	2,893	4,847
Not on list	3,975	124,042	20,153	0	148,170
Total	8,853	129,120	24,281	11,191	173,445

Heavily Hispanic names are those where more than 75% identify as Hispanic. Rarely Hispanic names are those where 5% or fewer identify as Hispanic. Table 3 shows that the lists are quite different – only 10,483 of the 173,445 names have the same designation in both tables (highlighted in green). However, these 10,483 names include nearly all of the most common names, so that a hypothetical cancer cohort with the same name distribution as the 2010 census would see crude Hispanic cancer rates drop by just 2%. In practice, the changes in crude rates are even smaller than this, because the surname portion of the algorithm is not applied to all cancer cases – in the most commonly used option, only to those reported by facilities as 7 (surname only) or 9 (unknown).

Results & Discussion - continued

Table 4. Comparison of 1990 and 2010 Hispanic surname lists

Registry	Change in crude Hispanic rate (2010 vs. 1990)
New York (using NY-NHIA)	0.1%
New York (using NHIA)	-0.2%
California	-0.2%
Wisconsin	-0.6%
Massachusetts	0.1%

Table 4 gives the changes in crude Hispanic rates between the 2010 and 1990 surname lists for New York, California, Wisconsin, and Massachusetts. Note that each of the comparisons were made with recent data, but each state chose different diagnosis years.

Table 5. Example surnames that went from heavily to rarely or vice versa

1990 Heavily to 2010 rarely	1990 rarely to 2010 heavily
BARETTA	DZIB
CADAVONA	FILPO
CONSENTINO	GIBOYEAUX
FALTERMEIER	JERONIMO
GELBMAN	MACEDONIO
GRAYTON	MEDAL
INSANA	PRUDENCIO
SULA	RAMIRE
TALADAY	WISCOVITCH
ZABOR	XICOTENCATL

A number of names changed from heavily to rarely Hispanic or vice versa (several examples are given in Table 5). In general, these are names with a sample size of 1 in 1990 and just above 100 in 2010. Some may not appear to resemble Spanish names, but Google searches confirm all of them.

Conclusions

- ➢ The differences between NHIA and NY-NHIA are small enough to not be of particular concern.
- ➢ Registries may wish to review their cases coded as "other specified Spanish/Hispanic origin" (code 5) to see if a more specific code is suggested by the birth country.
- ➢ The updated Hispanic surname list does not appear likely to trigger any dramatic changes in rates among NAACCR member registries.

Acknowledgements: This work was supported in part by the Centers for Disease Control and Prevention's Cooperative Agreement U58/DP003879, awarded to the New York State Department of Health through the National Program of Cancer Registries. It would not have been possible without the diligence of the dedicated Certified Tumor Registrars of the hospitals of New York or the coding of the New York State Cancer Registry.

Special thanks to Mary Mroszczyk of the Massachusetts Cancer Registry, Brenda Giddings of the California Cancer Reporting and Epidemiologic Surveillance Program (CalCARES) Program, and Laura Stephenson of the Wisconsin Cancer Reporting System for their help with testing the new surname list.

*The 2010 surname list includes all names occurring more than 100 times in the 2010 census, so it is not strictly population-based, but nearly so. See https://www.census.gov/topics/population/genealogy/data/2010_surnames.html