Variation in Hispanic Self-Identification, Spanish Surname, and Geocoding: Implications for Ethnicity Data Collection

Debra P. Ritzwoller*, Nikki Carroll, Bridget Gaglio, Anna Sukhanova, Fabio A. Almeida, Melanie A. Stopponi and Diego Osuna

Institute for Health Research Kaiser Permanente Colorado, Denver, CO 80237-8066, USA

Abstract: This study examines the variation in surname analysis and geocoding, and their association with self-identified Hispanics in an HMO. We collected ethnicity data from three studies, and employed Spanish surname software and census tract level geocoding to create proxies for Hispanic ethnicity. We computed sensitivity, specificity, and estimated multivariate logistic regression models to examine the variation in the likelihood of a match between self-identified Hispanics and surname. Sensitivity and specificity with respect to surname varied across the three studies, ranging from 57%-91% and 89%-96%, respectively. Relative to self-report, the sensitivity of the census tract measure of density of Hispanics, varied from 5%-15%. Multivariate models suggest that the likelihood of a match between self-identified Hispanics and surname was not associated with age or gender. Self-identified Hispanics living in neighborhoods with the highest density of Hispanics were less likely than those in more mixed neighborhoods to have a Spanish surname. Employing the Spanish surname software on only densely populated Hispanic census tracts may not always improve the likelihood of correctly identifying Hispanic subjects.

INTRODUCTION

In order to address racial and ethnic disparities in both health care and in the delivery and receipt of behavioral health interventions, researchers and policy makers need information regarding the racial and ethnic mix of their target populations. Several studies have raised concerns regarding the gaps in the collection of accurate racial and ethnic data at the population level [1-4]. Self-reported ethnic and racial classification is considered the "gold standard" compared to administrative methods [5]. In the absence of self-report data, research suggests that combined surname analysis and geocoding may provide an accurate and efficient means of inferring race/ethnicity across health plan membership or other populations [6].

Beginning in 1950, the United States Census Bureau produced and released a decadal Spanish Surname list. The basis for including a specific surname on that list was the similarity of that name's geographic distribution of the Hispanic origin population within the United States [7]. The 1970 census was the first that allowed the opportunity for people to self-identify as a "person of Spanish origin," which was later changed to "Hispanic."

The Generally Useful Ethnic Search System, or GUESS, is a previously validated surname program for identifying Hispanic subpopulations in the southwest [8-12]. The GUESS program was revised and validated using state tumor registry records by the Lovelace Clinic Foundation in Albuquerque, New Mexico [6]. Published sensitivities associated with the GUESS program and the US Census Spanish surname list range from 82-95% for males, and 67-82% for females [8, 10-12].

The methodology of geocoding residential addresses by linking them to U.S Census Bureau data and then using area-based socioeconomic measures in the study of health related outcomes has been extensively employed as a relatively inexpensive solution to the absence of population based, self-reported data socioeconomic status (SES) data including race/ethnicity, income and education [13, 14]. Residential addresses may be linked at different geographic levels that by size, and include zip code, census tract, and census block. Zip codes have an average of population of 30,000 and are considered administrative units established by the United States Postal Service. Census tracts on average contain 4,000 individuals and tend to be more homogeneous than zip codes. The smallest geographic unit for which census socioeconomic data are tabulated are block groups, which average approximately 1,000 individuals. Research results suggest that census tract may perform equally or better in detecting SES gradients in health than census block [15, 16]. The optimal "cut-point" (e.g., percent of residence by race or ethnicity) used to classify individuals has been suggested to be at least at 50 percent [6, 17].

In this study we use the "gold standard" of self-report to examine the variation in surname analysis and geocoding and examine the likelihood that geocoded measures will enhance the efficiency of surname identification of Hispanic or Latino members of a health maintenance organization (HMO), who were recruited for three behavioral health interventions.

MATERIALS AND METHODS

Setting and Study Populations

We collected self-identified race/ethnicity data from surveys administered during three behavioral interventions conducted at Kaiser Permanente Colorado (KPCO), during 2005

*Address correspondence to this author at the Institute for Health Research Kaiser Permanente Colorado, Denver, CO 80237-8066, USA;
Tel: 303- 614-1317; Fax: 303- 614-1305;
E-mail: debra.ritzwoller@kp.org or debra.ritzwoller@kpco-ihr.org

1874-9240/08

and 2006: a tailored smoking reduction project – *Smoking Less, Living More*, a lifestyle behavior intervention which targeted Latina women with type 2 diabetes at risk for CHD – *¡Viva Bien!*, and an Internet-mediated program to enhance consumption of fruits and vegetables –*Making Effective Nutritional Choices* or *MENU*. This study was approved by the KPCO Institutional Review Board.

KPCO is a closed panel, group model, non-profit HMO providing integrated health care services to 465,000 members (approximately 16% of the insured population) in the Denver-Boulder metropolitan area. At the time these studies were conducted, race and ethnicity data were not collected from all members. A variety of member surveys combined with census data suggest that the demographic characteristics of KPCO members highly correspond to the characteristics of the Denver metropolitan population. In comparison to the Denver metropolitan area, the KPCO population has a slight overrepresentation of females and an overrepresentation of individuals over age 50 years. During the study period, KPCO was one of the few Medicare managed care providers in the Denver-metropolitan area, therefore the membership is over represented by members over the age of 65. In addition, given that the majority (>65%) of KPCO population is comprised of a predominately employer based commercially insured members, the KPCO membership is underrepresented by individuals in the lowest income strata.

Smoking Less, Living More is a randomized controlled trial targeting smokers scheduled for an outpatient surgery or diagnostic procedure who are unwilling or unable to quit, but are willing to make an attempt at decreasing the number of cigarettes they smoke [18, 19]. Data from KPCO's electronic medical record were used to identify potential participants for the project. These data included administrative data from day surgery and diagnostic procedures scheduled (type of procedure and procedure date), smoking status, name, address, and age. A description of the project and an "opt-out" postcard were sent to subjects meeting the inclusion criteria three weeks before the procedure date. One to two weeks after the letter was sent, a trained interviewer called participants who did not decline to explain the study and determine if the person met eligibility requirements. In order to assess the representativeness of participants, baseline demographic, race/ethnicity, and smoking history were collected on all those contacted. These data included the question: "Do you consider yourself to be Hispanic or Latino?"

The *¡Viva Bien!* program is an adaptation for Latinas of previously successful research [20] targeting diet, physical activity, stress management, social support, and smoking cessation in women 40 years of age and older with type 2 diabetes. KPCO's EMR was used to identify potential subjects who had been diagnosed with type 2 diabetes for at least six months. The sample included in this analysis was limited to the west area of KPCO's Denver service area. The initial recruitment package included a cover letter briefly explaining the project and inviting potential subjects to participate, a brochure describing the study, and a stamped, self-addressed postcard to return to KPCO asking ethnicity and whether the person wished to receive further contact about the study. To determine ethnicity eligibility, the postcard asked, "Do you consider yourself Hispanic, Latina, or Mexican?"

The third sample included in this analysis is based on data collected as part of MENU, a project of the Cancer Research Network (http://crn.cancer.gov) [21]. KPCO is one of five CRN sites participating in MENU, an Internet-mediated program to enhance consumption of fruits and vegetables. Each site randomly selected 6,000 eligible patients, age 21-65 years, stratified by gender. In order to enhance minority participation at KPCO, the sample was further stratified by Hispanic indicator derived from surname analysis. Identified members received an introductory letter, a $2 enrollment incentive, a study description, and information on the MENU website. On the study website, participants could confirm their eligibility and complete registration with a baseline survey that asked, "Are you of Hispanic or Latino origin or descent?"

Surname Analysis

The surname of subjects from all three studies who were contacted and who provided consent for data collection, were mapped to the Lovelace modified GUESS program described above. Surname was captured through KPCO administrative databases. The data set containing the participant's last name was coded to prepare the file for comparison: a) embedded titles such as "Jr.," "Sr.," "III," etc. were removed, b) embedded blanks were removed so that De La Cruz became DELACRUZ, c) all names were placed in separate last name fields. Each last name field was compared with the GUESS program and a probable Hispanic indicator flag was created.

Geocoding

In order to capture socioeconomic proxy measures, we geocoded all three samples using MapMarker® Plus and 2000 Census data. Due to missing or incorrect details within each subject's recorded address, mapping each subject's address to the census block level resulted in a large number of missing observations; therefore, we used census tract to capture median household income and the proportion of the census tract self-identifying as Hispanic. We created six categories based on the proportion of Hispanic households in the census tract: $\geq$ 75 percent, 50-74 percent, 25-49 percent, 5-24 percent, and <5 percent. In addition, we created two dichotomous census tract indicators that we constructed to be consistent with prior literature that suggests one should employ a measure consistent with the majority population [6, 17]. We created a dichotomous variables denoting that > 50% of the census tract self-identified as Hispanic and a dichotomous variable denoting whether the median household income for the subject's census tract was at or above the 2000 U.S Census reported median income of $41,994.

Analyses

Descriptive statistics were estimated for the characteristics associated with each study population. We calculated mean age, and the distribution of gender and all census-derived variables. Using self-reported ethnicity as the gold standard, we computed sensitivity and specificity to assess the accuracy of the GUESS program. Sensitivity was defined as the percentage of self-identified Hispanics who were also classified as Hispanic by the GUESS algorithm, and specificity as the percentage of self-identified non-Hispanics who also were classified as non-Hispanic by the GUESS algo-

rithm. We also examined the distribution of self-identified Hispanics by categories of census tract Hispanic population density and estimated the sensitivity and specificity of Hispanic self-identification and the dichotomous census tract measure of >50 percent self-identified Hispanic.

To enhance sample size and variability in the demographic factors, we pooled data from all three studies and employed logistic regression to examine the potential association of the demographic factors with the likelihood of a match between self-identified Hispanic and the GUESS program. The sample used in this model was limited to subjects self-identifying as Hispanic whose address could be mapped to a valid census tract. Exclusions to latter category include participants reporting only a post office box or those reporting an incomplete or invalid address. Covariates in the model included a dichotomous variable noting the project the subject was enrolled in (*Smoking Less, Living More; ¡Viva Bien!;* or *MENU*), age, gender, and whether or not the subject lived in a census tract with reported median income at or above the national average. To test the hypotheses that a higher density of self-identified Hispanics within an individual's census tract will improve the likelihood of a match between self-report and the GUESS system, we included five dichotomous variables that denote the proportion of Hispanics in the individuals census track: < 5 percent (reference case), 5- 24 percent, 25-49, 50-74 percent, and >75 percent. All analyses were performed using SAS Software 9.1 (SAS Institute, Cary, NC).

RESULTS

Table **1** describes the variation across the three samples for measures of age, gender, proportion of self-identified Hispanics, proportion mapping to the revised GUESS program, census tract proxies of median household income, the

Table 1. Demographics, the Proportion of Self-Identified Hispanics, the Proportion Mapping to GUESS, and the Proportion Mapping to Census Tract Measures, for Subjects Enrolled in Three Behavioral Intervention Studies Conducted at Kaiser Permanente Colorado, 2005-2006

	Study Name		
	Smoking Less, Living More	***¡Viva Bien!***	**MENU**
Number of Observations	593	1,185	513
Mean Age (SD)	55.33 (11.23)	61.05 (8.97)	44.22 (11.19)
Percent Male (n)	29.17% (173)	0	39.97 (205)
Percent Hispanic from self-report (n)	7.42% (46)	14.93% (177)	28.46% (146)
Percent Hispanic self-report- male (n)	5.2% (9)	-	28.29% (58)
Percent Hispanic self-report – female (n)	8.33% (35)	-	28.57% (88)
Percent mapping to GUESS software (n)	7.76% (46)	15.86% (188)	34.89 % (179)
Percent mapping to GUESS software – male (n)	4.05% (7)	-	31.22% (64)
Percent mapping to GUESS software – female (n)	9.29% (39)	-	37.34% (115)
Census Tract - % Population > Median Income	63.41% (345)	55.57% (554)	77.03% (359)
Census Tract: > 50% Hispanic Population for all subjects (n)	4.73% (28)	10.30% (122)	2.53% (13)
Census Tract - % Hispanic Population for all subjects (n)			
Less than 5%	33.56% (199)	33.92% (402)	36.26% (186)
5 – 24%	39.80% (236)	30.80% (365)	41.33% (212)
25 – 49%	13.66% (81)	10.04% (119)	10.72% (55)
50 – 74%	3.88% (23)	5.74% (68)	1.36% (7)
Greater than 75%	0.85% (5)	4.56% (54)	1.17% (6)
Missing	8.26% (49)	14.94 (177)	9.16% (47)
Census Tract: > 50% Hispanic Population for all subjects self-identifying Hispanic (n)	15.91%(7)	40.11 (71)	4.79% (7)
Census Tract - % Hispanic Population for all subjects self-identifying Hispanic (n)			
Less than 5%	11.36% (5)	13.56 % (24)	27.40% (40)
5 – 24%	34.09% (15)	20.34% (36)	42.47% (62)
25 – 49%	29.55% (13)	16.38% (29)	17.81% (26)
50 – 74%	11.36% (5)	19.21% (34)	1.37% (2)
Greater than 75%	4.55% (2)	20.90% (37)	3.42% (5)
Missing	9.09% (4)	9.60% (17)	7.5% (11)

distribution of all subjects census tract measure of Hispanic population density, and the distribution of self-identified Hispanics census tract measure of Hispanic population density. Table **2** describes the measures of sensitivity and specificity across the three studies.

Table 2. Sensitivity and Specificity of Hispanic Self Report to GUESS Surname, and Hispanic Self Report to Geocoded Census Tract Measure Tract of Greater than a 50 Percent Hispanic Population

	Study		
	Smoking Less, Living More	***¡Viva Bien!***	**MENU**
N (from Self-report)	593	1,185	513
Hispanic Self-Identification *vs* GUESS			
Sensitivity	56.82%	78.53%	91.10%
Specificity	96.17%	95.14%	87.47%
Sensitivity by Gender			
male	66.67%		94.83%
female	54.29%	78.53%	88.64%
Specificity by Gender			
male	99.39%		93.88%
female	94.81%	95.14%	83.18%
Hispanic Self-Identification *vs* Geocode			
Sensitivity	15.91%	40.11%	4.79%
Specificity	96.17%	94.94%	100%
Sensitivity by Gender			
male	0		3.45%
female	20.00%	40.11%	5.68%
Specificity by Gender			
male	93.90%		97.28%
female	97.14%	94.94%	99.09

The average age of the three samples varied from the eldest, *¡Viva Bien!*, with a mean age of 61 years, to MENU, the youngest, at 44 years old. As described above, the entire *¡Viva Bien!* sample was female, while MENU 40% male and Smoking Less, Living More was approximate 29% male.

Of the 593 subjects contacted regarding the Smoking Less, Living More study who responded to the Hispanic self-identification question, 44 (7.42%) self-identified as Hispanics. The surnames of 46 subjects (7.76%) mapped to the GUESS software. However, only 25 self-identified Hispanic mapped to the surname software, for a sensitivity of 57% and a specificity of 96%. Less than 5 % of the overall sample lived in census tract reporting > 50% Hispanic and only 15% of those self-identified as Hispanics lived in a census tract reporting a Hispanic population > 50%.

Of the 1,185 females who were contacted for the *¡Viva Bien!* study and who provided their race/ethnicity, 177 (14.93%) self-identified as Hispanic. The surnames of 188 subjects (15.86%) mapped to the GUESS software, while 139 of the 177 self-identified Hispanic or females mapped to the GUESS software, for a sensitivity of 79% and a specificity of 95%. Almost 10% of the *¡Viva Bien!* sample lived in census tracts reporting >50% Hispanic, and 40% of self-identified Hispanics lived in a census tract with a Hispanic population >50%.

Of the 513 MENU subjects who were contacted and responded to the Hispanic identification question, 146 (28.46%) self-identified as Hispanic. The surnames of 179 subjects (34.89%) mapped to the GUESS software, while 122 of the 179 self-identified Hispanics mapped to the GUESS software for a sensitivity of 91% and a specificity of 89%. Less than 3% of the full sample lived in a census tract with a Hispanic population >50%, and <5% of self-identified Hispanics lived in a census tract with a Hispanic population >50%.

Table **2** also describes the variation by gender for the two measures of sensitivity and specificity (self-report *vs* GUESS, and self-report *vs* geocoding) across the three studies. For females, the sensitivity of the GUESS system varied from 54 percent in Smoking Less, Living More to almost 89 percent for MENU; sensitivity for females was lower than for males in both studies (*¡Viva Bien!* was limited to females). The GUESS sensitivity gender differential between the two studies was 12.38 percent Smoking Less, Living More and 6.19 percent for MENU. The GUESS specificity was the lowest for MENU, overall, and for both males and females. The gender variation associated with the sensitivity of self-report and the geocoded census tract measure of greater than 50 percent Hispanic varied from 0 for males to 20 percent in females in Smoking Less, Living More. In MENU, the sensitivity was 5.68 percent, about 2 percent higher than for males.

Results from the logistic regression model are described in Table **3**. Consistent with descriptive statistics in Tables **1** and **2**, the surname of self-identified subjects in the MENU and the *¡Viva Bien!* projects were significantly ($p < 0.03$) more likely to map to the revised GUESS program than those subjects enrolled in Smoking Less, Living More. However, when other characteristics were adjusted for, the mean percentages of matches for each study did not significantly change from those described in Table **1**.

Controlling for age, gender, and census tract median income flag, three of the census categories denoting the proportion of Hispanic population were significantly associated with the likelihood of a match ($p < 0.03$). Subjects living in census tracts with 25% - 49% Hispanic representation were seven times more likely ($p < 0.001$) to have a match than those living in tracts with less than 5% Hispanic representation. However, as the concentration of Hispanic representation increased, the likelihood of a match decreased, but remained significant (to the reference case of < 5%). Specifically, the size of the coefficient and the associated odds-ratios decreased relative to the 25-49% category.

DISCUSSION

Using convenience samples derived from three very different behavioral interventions that were conducted in an HMO located in a large metropolitan area, we found that the

Table 3. Multivariate Logistic Regression Model: Likelihood of Self-Identified Hispanics Mapping to Surname Program. N= 335

Variable	Estimate	Std. Error	Odds Ratio	95% Confidence Limits	p-Value
Intercept	1.63	2.43			0.50
Smoking Less, Living More	Ref.				
MENU	2.44	0.51	11.51	4.23, 31.39	<.0001
Viva Bien	1.00	0.44	2.72	1.14, 6.48	0.02
Age	-0.10	0.09	0.91	0.76, 1.08	0.28
Age^2	0.001	0.001	1.00	0.99, 1.00	0.20
Gender (male = 1)	-0.59	0.54	0.55	0.19, 1.59	0.27
Census Tract < 5% Hispanic	Ref.				
Census Tract 5-24% Hispanic	0.52	0.40	1.68	0.77, 3.65	0.19
Census Tract 25-49%Hispanic	1.96	0.59	7.08	2.21, 22.63	0.001
Census Tract 50-75% Hispanic	1.63	0.63	5.08	1.50, 17.23	0.009
Census Tract >75% Hispanic	1.36	0.60	3.90	1.20, 12.65	0.024
Income >= U.S. Median	0.14	0.39	1.15	0.54, 2.46	0.713

-2 Log Likelihood for intercept at covariates = 320.49, Likelihood Ratio under the Chi-Square = 37.17, with 7 degrees of freedom (p = < 0.0001).

sensitivity of the GUESS program varied from 57-91 percent overall, 54-88 percent for females and 67 and 95 percent for males (two of three studies). The sensitivity of the geocoded census tract measures were low, and varied from approximately 5-40 percent. In this insured/employed and/or Medicare aged population, we found that subjects residing in heavily Hispanic neighborhoods were less likely than those in a neighborhood of <49% Hispanics, to be associated with of a match between self-identified Hispanics and the GUESS program.

With the exception of the estimated sensitivity of 54 percent derived from our small sample of self-identified Hispanic female subjects in Smoking Less, Living More (n=35), our findings are consistent with other published estimates of the sensitivity and specificity of Spanish surnames for women compared to self-reported ethnicity of 67 percent and 88 percent [8-10, 12]. A recent study by Sweeny and colleagues [11], estimated the sensitivity of GUESS or Spanish Surname lists at 71.6 percent for cases, representing women with breast cancer living in New Mexico or Utah, and 59.8 percent for the controls.

A study of a Kaiser Permanente population in the San Francisco area [9] reported that compared with self-identified ethnicity, a Spanish surname was 88.4 percent sensitive in identifying Latino men and 70.4 percent sensitive in identifying Latina women. In our study, the gender differential in sensitivity only reached 12 percent in the Smoking Less, Living More study where only 9 of 173 males self-identified as Hispanic. Spanish surname analyses both falsely identifies a large number of non-Hispanic persons as Hispanic and failed to identify a small proportion of Hispanics. In the San Francisco study [10], a large proportion of those falsely identified as Hispanic were non-Hispanic Filipinos with Spanish surnames. This points to the geographic variability in the accuracy of surname analysis [10, 22]. Surname identification performs best in areas with established and high density Hispanic populations [7].

We do not have information from our subjects regarding the source of the mis-identification in this analysis. However, prior research has demonstrated that English-speaking married and previously married women and those with higher socioeconomic status showed the highest discordance between self-reported ethnicity and surname coding [23]. The *¡Viva Bien!* was limited to females, and all three samples were derived from an insured population. Those in the MENU project had to have access to the Internet, and more than half of respondents from all three studies lived in a census tract where the reported median household income was above the 2000 national median household income of $41,994. Therefore, it is probable that these insured subjects in this study may represent a more acculturated and privileged cohort than the overall Hispanic population.

The MENU project had the highest concordance between self-reported Hispanic ethnicity and Spanish surname overall and for females, but the lowest number of self-identified Hispanics living in a census tract with a concentration of Hispanics greater than 75% (3.42%). This is consistent with other published studies that note that Hispanics live in far less segregated neighborhoods than African Americans [22, 23]. We also found a non-linear relationship between the likelihood of a match between a Spanish surname and increasing density of a Hispanic census tract. This finding is consistent with Logan *et al.* (2001) [24, 25], who estimates that the typical Hispanic lives in a neighborhood that is 45.5% Hispanic. Consistent with Chen *et al.* (2004), the geocoded measures that we used for identifying Hispanic members may be less accurate in more acculturated, homogeneous, and higher socioeconomic population like the cohort used in this study [26].

This study has several limitations. The study cohort was derived from three behavioral interventions that do not represent a random sample of the KPCO population. Identification and recruitment of subjects varied significantly across the three projects. Two of the interventions actively at-

tempted to recruit a high proportion of Hispanics. Initial recruitment for *¡Viva Bien!* targeted the western suburbs of Denver, which are known to be more heavily Hispanic. Approximately one-half of the MENU project's initial mailing was sent to members identified as having a Spanish surname. In addition, the sample of self-identified Hispanics from Smoking Less, Living More was very small, which limits our ability to generalize across the range of sensitivities that we found. We do not have information from the subjects to determine the sources of discordance.

During the recruitment period of these studies, no race/ethnic data were available. However, in order to promote the delivery of culturally and linguistic appropriate services at KPCO, a significant effort is currently underway to collect self-reported measures of race, ethnicity and language preferences. However, these efforts are not without barriers. A recent study by Baker *et al.* (2005) found that 80% of patients agreed that health care providers should collect information regarding race and ethnicity, but many felt uncomfortable giving this information [27]. Additionally, in the fall of 2006, two weeks prior to a KPCO mailing that included a survey asking for self-report of race and ethnicity, there was significant media coverage of Immigration and Customs Enforcement raids of two large Denver construction sites that took place where more that 100 undocumented workers were detained [28]. Until complete data is available, health plans and other public health entities may need to rely on ethnicity estimates derived from surname analysis and geocoding to meet HEDIS requirements, applications for federal grants, and other reporting needs.

CONCLUSION

We found significant variation in the sensitivity or concordance between self-identified Hispanics and Spanish surname and geocoded proxies. Fincella and Fremont (2006) note that the advantages and limitation of geocoding and surname analysis may complement each other and provide an attractive means of inferring race/ethnicity among health plan members [6]. Perhaps, as suggested by Fincella and Fremont, geocoded data could be used to generate a priori probabilities before assigning ethnicity based on surnames. Until such hypothesis are confirmed and these newer methods are employed, our findings suggest that significant care must be taken if these surname or geocoded proxies are used to assess quality of care, to assess disparities in care, or to assign biologic, behavioral or demographic characteristics to populations [29].

However, surname analysis can be useful in a variety of situations such as recruitment of minority patients for studies and for planning services aimed at specific populations. While the GUESS software is generally useful in identifying Hispanics *vs* non-Hispanics whites, it would not be useful in identifying the racial or ethnic identity where surname is not reflective of race/ethnicity (e.g., African-Americans). Thus, one must use it with care in any population with significant percentage of African-Americans, since it would classify African-Americans with non-Hispanic whites.

ACKNOWLEDGEMENTS

This research was supported by National Cancer Institute grants 01 CA 90974-01 and U19 CA079689, and National Heart Lung and Blood Institute grant R01 HL077120. We would like to thank all of our colleagues associated with the Smoking Less, Living More, *¡Viva Bien!*, and MENU projects. We would also like to thank Elizabeth Staton for her assistance in editing the original manuscript.

ABBREVIATIONS

GUESS = Generally Useful Ethnic Search System

HEDIS = Health Plan Employer Data and Information Set

HMO = Health Maintenance Organization

KPCO = Kaiser Permanente Colorado

SES = Socioeconomic status

REFERENCES

[1] Nerenz DR, Hunt KA, Escarce JJ. Health care organizations' use of data on race/ethnicity to address disparities in health care. Health Serv Res 2006; 41(4 Pt 1): 1444-50.

[2] Moscou S, Anderson MR, Kaplan JB, Valencia L. Validity of racial/ethnic classifications in medical records data: An exploratory study. Am J public Health 2003; 93(7): 1084-86.

[3] Bierman AS, Lurie N, Collins KS, Eisenberg JM. Addressing racial and ethnic barriers to effective health care: The need for better data. Health Aff (Millwood) 2002; 21(3): 91-102.

[4] Moy E, Arispe IE, Holmes JS, Andrews RM. Preparing the national healthcare disparities report: Gaps in data for assessing racial, ethnic, and socioeconomic disparities in health care. Med Care 2005; 43(3 Suppl): I9-16.

[5] Arday SL, Arday DR, Monroe S, Zhang J. HCFA's racial and ethnic data: Current accuracy and recent improvements. Health Care Financ Rev 2000; 21(4): 107-16.

[6] Fiscella K, Fremont AM. Use of geocoding and surname analysis to estimate race and ethnicity. Health Serv Res 2006; 41(4 Pt 1): 1482-500.

[7] Perkins RC. Technical Working Paper 4. Evaluating the Passel-Word Spanish surname list 1990 decennial census post enumeration survey results. Washington DC, US Bureau of the Census, Population Division. August 1993. [Accessed May 4, 2007] Available from: http://www.census. gov/population/www/techpap.html

[8] Howard CA, Samet JM, Buechley RW, Schrag Sd, Key CR. Survey research in New Mexico Hispanics: Some methodological issues. Am J Epidemiology 1983; 117(1): 27-34

[9] Rosenwaike I, Bradshaw BS. The status of death statistics from the Hispanic popupulation in the Southwest. Soc Sci Q 1988; 69(3): 722-36.

[10] Perez-Stable EJ, Hiatt RA, Sabogal F, Otero-Sabogal R. Use of Spanish surnames to identify Latinos: Comparison to self-identifciation. J Natl Cancer Inst Monogr 1995; 18: 11-5.

[11] Morgan RO, Wei II, Virnig BA. Improving Identification of Hispanic males in medicare: Use of surname matching. Med Care 2004; 42(8): 810-16.

[12] Sweeney C, Edwards SL, Baumgartner KB, *et al.* Recruiting Hispanic women for a population-based study: Validity of surname search and characteristics of nonparticipants. Am J Epidemiol 2007; 166(10): 1210-19.

[13] Krieger N. Overcoming the absence of socioeconomic data in medical records: Validation and application of a census-based methodology. Am J Public Health 1992; 82(5): 703-10.

[14] Pearl M, Braveman P, Abrams B. The relationship of neighborhood socioeconomic characteristics to birthweight among 5 ethnic groups in California. Am J Public Health 2001; 91(11): 1808-14.

[15] Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: The public health disparities geocoding project. Am J Public Health 2005; 95(2): 312-23.

[16] Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: A comparison of area-based socioeconomic measures--the public health disparities geocoding project. Am J Public Health 2003; 93(10): 1655-71.

[17] Kwok RK, Yankaskas BC. The use of census data for determining race and education as SES indicators: A validation study. Ann Epidemiol 2001; 11(3): 171-7.

[18] Levinson AH, Glasgow RE, Gaglio B, Smith TL, Cahoon J, Marcus AC. Tailored behavioral support for smoking reduction: Development and pilot results of an innovative intervention. Health Educ Res 2008; 23(2): 335-46.

[19] Glasgow RE, Gaglio B, France EK, *et al*. Do behavioral smoking reduction approaches reach more or different smokers? Two studies: similar answers. Addict Behav 2006; 31: 509-18.

[20] Toobert DJ, Strycker LA, Glasgow RE, Barrera M Jr, Angell K. Effects of Mediterranean lifestyle trial on multiple risk behaviors and psychosocial outcomes among women at risk for heart disease. Ann of Behav Med 2005; 29: 128-37.

[21] Wagner EH, Greene SM, Hart G, *et al*. Building a research consortium of large health systems: The cancer research network. J Natl Cancer Inst Monogr 2005; 35: 3-11.

[22] Hazuda HP, Comeaux PJ, Stern MP, Haffner SM, Eifler CW, Rosenthal M. A comparison of three indicators for identifying Mexican Americans in epidemiologic research. Methodological findings from the San Antonio heart study. Am J Epidemiol 1986; 123(1): 96-112.

[23] Winkleby MA, Rockhill B. Comparability of self-reported hispanic ethnicicty and Spanish surname coding. Hisp J Behav Sci 1992; 14: 487-95.

[24] Logan J. Ethnic diversity grows, neighborhood integration lags behind. Albany, NY: Lewis Mumford Center, University at Albany 2001.

[25] Logan J. Hispanic populations and their residential patterns in the metropolis. Albany, NY: Lewis Mumford Center, University at Albany 2002.

[26] Chen W, Petitti DB, Enger S. Limitations and potential uses of census-based data on ethnicity in a diverse community. Ann Epidemiol 2004; 14(5): 339-45.

[27] Baker DW, Cameron KA, Feinglass J, *et al*. Patients' attitudes toward health care providers collecting information about their race and ethnicity. J Gen Intern Med 2005; 20(10): 895-900.

[28] Quintero F, Ramirez R, Frank L. "Dozens nabbed on Buckley job" Rocky mountain news, September21, 2006. [Accessed May 15, 2007] Available from: http://www.rockymo untainnews.com/drmn/local/article/0,1299,DRMN_15_5379070,00. html

[29] "Federal data on race and ethnicity abound, but need to be interpreted locally to be useful in reducing health care disparities." Grant Results. Robert Wood Johnson Foundation. [Accessed March 15, 2008] Available from: http:// www.rwjf.org/programareas/resources/grants-report.jsp?filename=050333.htm&pid=1142&gsa=1

Received: May 25, 2008 Revised: July 16, 2008 Accepted: July 21, 2008