When Race and Hispanic Origin Reporting are Discrepant Across Administrative Records Sources: Exploring Methods to Assign Responses

Sharon R. Ennis, Sonya Rastogi, and James Noon
U.S. Census Bureau

DRAFT: DO NOT CITE OR QUOTE WITHOUT PERMISSION

Disclaimer: This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

ABSTRACT

The U.S. Census Bureau is researching uses of administrative records in survey and decennial census operations. One potential use of administrative records is to utilize the data when race and Hispanic origin responses are missing. When federal and third party administrative records are compiled, race and Hispanic origin responses are not always the same for an individual across different administrative records sources. We explore different sets of business rules to assign one race and one Hispanic response when these responses are discrepant across sources. We also describe the characteristics of individuals with matching, non-matching, and missing race and Hispanic origin data across several demographic, household, and contextual variables. We find that minorities, especially Hispanics, are more likely to have non-matching Hispanic origin and race responses in administrative records compared to the 2010 Census. Minority groups are also more likely to have missing race or Hispanic origin data in administrative records.

INTRODUCTION

The U.S. Census Bureau is researching uses of administrative records (AR) in survey and decennial operations in order to reduce costs and respondent burden while preserving data quality. One potential application of administrative records is to utilize the data when race and Hispanic origin responses are missing.

Race and Hispanic origin data collected by the Census Bureau are critical to the legislative redistricting process and in enforcing compliance with the Civil Rights Act, Voting Rights Act, Fair Housing Act, and Equal Employment Opportunity Act. These data are also used by federal programs, researchers, and policymakers in assessing racial and ethnic disparities in health, employment, income, and housing, for example.

Item nonresponse for race and Hispanic origin is relatively low. However, when a respondent does not provide a race or Hispanic origin, the Census Bureau employs methods such as hot decks to impute a response. A hot deck is geographically based, where responses from a nearest neighbor are used to impute missing responses to people with similar characteristics. The underlying assumption of a nearest neighbor hot deck is that people who live near each other share similar characteristics; however, with increasing racial and ethnic diversity in the U.S., this is less likely to be true (Farber et al. 2005).

Administrative records may be used to improve imputation and to assign data to missing responses. For the first time in the 2010 Census, information that people had already provided in either Census 2000 or the 2001-2009 American Community Surveys were sometimes used in imputing missing race and Hispanic origin responses. In fact, previous census responses were used in almost 40 percent of all imputed Hispanic origin responses and 30 percent of all imputed race responses (Rothhaas et al. 2012). This suggests that this new imputation method could be valuable in assigning missing race and Hispanic origin data. We may be able to expand on this imputation method and include other federal and third party sources of administrative records.

However, when federal and third party administrative records are compiled, race and Hispanic origin responses are not always the same for an individual across different administrative records sources. In this paper, we explore different sets of business rules used to assign a single race and Hispanic origin response and evaluate which set of rules results in the highest level of agreement between the administrative records composite and the 2010 Census. We also describe the characteristics of individuals whose race or Hispanic origin responses in the administrative data match or do not match 2010 Census data, or have missing race or Hispanic origin responses in administrative records.

In the next sections of this paper, we provide background on previous research on race and ethnicity data in administrative records. Then we discuss the data and methods used in our analysis and present the results from our study. We conclude with a summary of our findings and propose future research.

BACKGROUND

Census Bureau Research on the Quality of Race and Hispanic Origin in Administrative Records

In response to expanding interest in the use of administrative records in enhancing a decennial census, the Census Bureau developed the Statistical Administrative Records System (StARS) in 1999. StARS 1999 was built from seven administrative files – the Internal Revenue Service (IRS) Individual Income Returns, IRS Information Returns, Department of Housing and Urban Development (HUD) Tenant Rental Assistance Certification System (TRACS), the Selective Service System Registration System, Indian Health Service (IHS) file, and the Center for Medicare and Medicaid Services (CMS) Medicare Enrollment Database (MEDB), and the Social Security Administration Numerical Identification File (Numident) (Farber and Leggieri 2002). StARS 1999 was developed to support the Administrative Records Experiment which simulated Census 2000 counts with administrative records (Farber and Leggieri 2002, Bye and Judson 2004). This previous research found that StARS had a lower representation of the minority population compared to Census 2000. However, one of the limitations of the administrative data in the StARS database was the inconsistent collection of race and ethnicity data. In particular, Numident, which provided the widest coverage of race and ethnicity for the population, included Hispanic as a race category and did not collect multiple race responses as done in Census 2000 (Farber and Leggieri 2002). To account for these limitations, StARS included modeled data used to impute race data for individuals who were missing race.

In a more recent study, the 2010 Census Match Study, Rastogi and O'Hara (2012) expanded on this research and evaluated the quality of demographic responses in administrative records, as defined by agreement between administrative records and the 2010 Census. In addition to the administrative sources used in StARS, this study utilized thirteen additional federal and third party files. Rastogi and O'Hara (2012) found that the quality of data in administrative records for persons identified as non-Hispanic in the 2010 Census was considerably higher compared to Hispanics. Results of the quality analysis of race data varied by race group. The White alone, Black alone, and Asian alone populations had higher quality race data in administrative records compared to the Two or More Races, Native Hawaiian or Other Pacific Islander (NHPI) alone, American Indian or Alaska Native (AIAN) alone, and Some Other Race (SOR) alone populations. In a study that replicated the 2010 Census Match Study using data from the 2010 American Community Survey, Bhaskar et al. (2014) found results for race and Hispanic origin that were consistent with those found by Rastogi and O'Hara (2012).

After Census 2000, research was also conducted to identify alternative imputation methods using administrative records. Farber et al. (2005) simulated imputation using administrative records and evaluated accuracy of the results. The agreement rates for race and Hispanic origin were 96 percent and 98 percent, respectively (Farber 2005). Obenski et al. (2005) also simulated imputation, comparing hot deck imputation to direct assignment methods using administrative records. Four states[1] were used in their analysis and they found that the administrative records method was more accurate than the hot deck method for Hispanic origin and race (Obenski et al. 2005). However, the authors did recommend that the administrative records assignment be followed by hot deck imputation for individuals who are not linked to administrative records.

Building on this previous research, Rastogi et al. (2014) found administrative records provide coverage of 13 percent of race responses imputed by hot decks in the 2010 Census. Three different surname-assisted hot decks (Spanish surname-assisted, non-Spanish surname-assisted, and non-surname-assisted) were used to impute Hispanic origin in 2010. The administrative records coverage rates were 18 percent for the non-surname hot deck imputed responses, and 20 percent and 42 percent, respectively, for the Spanish surname and non-Spanish surname hot deck imputed responses (Rastogi et al. 2014). Comparing 2010 Census hot deck-imputed race and Hispanic origin data to administrative records, the authors found that the agreement rate for non-Hispanic and Hispanic responses was 96 percent and 54 percent, respectively. The agreement rate between administrative records and hot deck-imputed responses for White alone was relatively high at 83 percent and considerably lower for Black alone at 52 percent. Agreement rates for the other race groups ranged from 1 percent for the multiracial population to 17 percent for the SOR alone population. The authors concluded that administrative records can improve the quality of imputations for missing race and Hispanic origin responses; however, hot deck imputation will still be necessary for individuals who are missing data in administrative records or who can not be linked to administrative records.

[1] The four states were Delaware, Georgia, New York, and Florida.

In addition to research on the quality of race and Hispanic origin responses in administrative records, previous research has been conducted on the fluidity of racial and ethnic responses and the characteristics of individuals with non-matching and missing race and ethnicity data.

Racial and Ethnic Fluidity

One reason an individual's race or Hispanic origin in administrative records may not match their response in census data is racial and Hispanic origin fluidity. There is a rich literature in sociological research that reports that individuals may change their identity over time or in different situations and contexts. Changes in race response vary by race group. Non-Hispanic Whites, Blacks, and Asians are usually consistent in their race responses; contrarily, race response change is more common among non-Hispanic AIAN, NHPI, and multiracial individuals (Liebler et al. 2014; Doyle and Kao 2007; del Pinal and Schmidley 2005; Bentley et al. 2003). Previous Census Bureau research from the 1990, 2000, and 2010 censuses shows that individuals are relatively consistent in their responses to the Hispanic origin question with three percent or less changing their answer between the census and its corresponding reinterview (Dusch and Meier 2012; Singer and Ennis 2003; U.S. Census Bureau 1993). However, other studies of adolescents have found that Hispanic students are relatively inconsistent in their ethnic identification (Perez 2008; Brown et al. 2006; Eschbach and Gomez 1998).

Prior research shows substantial racial fluidity among Hispanics relative to non-Hispanics (Liebler et al. 2014; Dusch and Meier 2012; Brown et al. 2006; Singer and Ennis 2003). One factor that may affect race reporting among Hispanics is that although the federal government defines race and ethnicity as separate concepts, many Hispanics view race and ethnicity as one concept and identify their race as "Hispanic." When faced with the federal standard racial categories, people who view their race as Hispanic may 1) not answer the race question, 2) report Hispanic responses that are tabulated as SOR, or 3) report a category that they feel may not be the best fit for their racial identity. Another factor affecting Hispanic racial identification is differences in questionnaire design. Campbell and Rogalin (2006) conducted a study that compared responses from separate ethnicity and race questions to a combined ethnicity and race question for the same respondent. The authors found that most Hispanics who chose a race in the separate question identified as Hispanic only to the combined ethnicity and race question.

Characteristics of People with Non-Matching and Missing Hispanic origin and Race Responses

Previous research that measured agreement of race and ethnicity data in administrative records with survey data has found that agreement varies by demographic and socioeconomic characteristics. Specifically, American Indians, Asians, Pacific Islanders, and Hispanics are more likely to have non-matching responses than Whites, Blacks, and non-Hispanics (Zaslavsky et al. 2012; Gomez et al. 2005; Kressin et al. 2003; Arday et al. 2000). Younger individuals and males are associated with inconsistent race and Hispanic origin responses (McAlpine et al. 2007; Gomez et al. 2005). Recent research examining the agreement between Medicaid administrative records and census data found that minorities, individuals that are older, male, and live in affluent neighborhoods are more likely to have non-matching race and Hispanic origin responses

(Fernandez et al. 2015). In a study comparing race and ethnicity data in Veteran Affairs administrative data and survey data, Kressin et al. (2003) find that patients whose race and ethnicity were consistently reported were more likely to live alone. However, when the authors excluded patients with missing race/ethnicity data in the administrative records source, those more likely to show agreement did not live alone.

In a study of Hispanic origin and race response change between Census 2000 and the 2010 Census, Liebler et al. (2014) found that Hispanics, American Indians, Pacific Islanders, multiracial people, children, people who reside in the West, and people who responded to either census via a mode other than mail were more likely to have non-matching race and ethnic responses.

Previous studies comparing survey data to administrative records found that White and younger individuals are more likely to have missing race responses (McAlpine et al. 2007; Kressin et al. 2003). However, Fernandez et al. (2015) found that individuals who are Hispanic, AIAN, older, male, and live in neighborhoods with higher median household incomes have a higher likelihood of having missing race responses in Medicaid administrative records. Women, minorities except Asian/NHPI, and those living in neighborhoods with higher median household incomes are more likely to have missing Hispanic origin responses in administrative records (Fernandez 2015).

DATA AND METHODS

The data in this study include federal and third party files used to build and assign demographic data to an administrative records composite. Federal data files used to assign demographic information include Previous Census Records (Census 2000 and American Community Survey 2001 to 2009), the Social Security Administration Numerical Identification File (Numident), three files from the Department of Housing and Urban Development (HUD), the Center for Medicare and Medicaid Services Medicare Enrollment Database (MEDB), the Indian Health Service (IHS) Patient Registration System, Temporary Assistance for Needy Families (TANF), a Texas Supplemental Nutrition Assistance Program file (SNAP), and the Medicaid Statistical Information System (MSIS). In addition, we used four third party files to assign demographic information. See Rastogi and O’Hara (2012) for more information on the data.

Administrative records sources vary in the collection of Hispanic origin and race data. Many of the federal files report race and ethnicity according to the Office of Management and Budget's (OMB) revised 1997 race and ethnic standards.[2] However, there are a couple of exceptions. The Numident and MEDB files treat race and Hispanic origin as one concept and have one combined race and ethnicity variable. In other words, the categories of the variable include "Hispanic" in addition to the race groups. Additionally, the Numident and MEDB data have a combined

[2] Federal agencies must adhere to race and ethnicity standards issued by the OMB. There are a minimum of two ethnicities: Hispanic or Latino and Not Hispanic or Latino. There are five categories on race: White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander. For respondents who do not identify with any of these five race categories, OMB approved the Census Bureau’s inclusion of a sixth category, Some Other Race. Respondents are also permitted to identify with more than one race. The standards are available online at <www.whitehouse.gov/omb/fedreg/1997standards.html>.

category for Asian and Pacific Islander and do not collect multiple responses or include a category for multiracial persons.

In order to compare the race and ethnicity data from the Numident and MEDB files to the 2010 Census, we had to recode the combined race and ethnicity variable into two separate variables, one for ethnicity and one for race. Individuals who were identified as Hispanic were coded as such with missing race information since we have no information about their race. Similarly, individuals who were identified as a race were coded as that race group with missing Hispanic origin information. For example, if an individual identified as Black, then the separate ethnicity variable was coded as missing and the race variable was coded as "Black."

Although the HUD and TANF files do collect race and Hispanic origin according to the OMB standard, these files do not include a category for SOR, unlike the Census Bureau. The IHS file only identifies individuals as either AIAN or non-AIAN. The third party files model race and Hispanic origin data using information on surname and geography.

The U.S. Census Bureau has linked individuals' census records as part of an effort to understand response variability and reduce data collection costs. We use internal Census Bureau data from the 2010 Census and link this with the administrative records composite. All person records were processed through the Person Identification Validation System (PVS), which used probability record linkage techniques and personal information such as name and date of birth to assign an anonymized Protected Identification Key (PIK) to each person, as possible, allowing record linkage across datasets (see Wagner and Layne 2014). The data do not include people who do not have a Social Security Number and those whose personal information was too ambiguous or incomplete to assign a PIK. Once the PIK was assigned in each separate data set, it was used to link a person's record in the 2010 Census to his or her own record in the administrative records composite.

The descriptive analysis of response matching between Hispanic origin and race in the administrative records composite and the 2010 Census is based on a link between the two files where the census information is unedited. These match rates are dependent upon the presence of Hispanic origin and race data assigned to the administrative record. Records without any available Hispanic origin or race data are not included in the descriptive match rates.

We perform multinomial regression analysis separately for Hispanic origin and for race. These models predict whether a linked Census-AR record matches on Hispanic origin or race (coded as "0"), whether the Hispanic origin or race data do not match (coded as "1"), and whether the AR record does not have any available Hispanic origin or race data (coded as "2"). Because the dependent variables include AR records with missing demographic data, the distributions for the dependent variables differ from the distribution for matching Hispanic origin and race data presented in the descriptive analysis. As with the descriptive statistics, the models are limited to census records that are unedited. Because all administrative records sources do not have an SOR category or collect multiple races, we excluded these groups from the regression analysis.

The independent variables for these regressions include individual-level demographic variables, household-level characteristics, tract-level contextual characteristics, and geographic region. Individual-level variables include the person's Hispanic origin, race, age, and gender as reported in the Census. Household-level variables include the household tenure (owner, renter, renter with no rent paid), household type and size as reported in the Census, the Census mode by which the household responded (mailout/mailback, nonresponse follow-up, or another mode), and whether the household is in an urban or rural area. In addition, tract-level variables measure the percent of non-Hispanic whites in the tract in the Census and the logged median household income in the tract according to the American Community Survey.

Limitations

Our analysis does not include people in administrative records who were not assigned a PIK. Bias may be introduced into our results if characteristics of individuals who received a PIK are different from those that did not receive a PIK. In addition, people in administrative records that did receive a PIK but could not be linked to 2010 Census data are not included in the analysis. This too is likely to result in some bias in our findings. Therefore, our results should be interpreted with caution.

PRELIMINARY FINDINGS

The final version of this paper will be a Center for Administrative Records Research and Applications Working Paper (which will be available at https://www.census.gov/srd/carra/) and will have a complete Results section and Conclusion section.

We find that minorities, especially Hispanics, are more likely to have non-matching Hispanic origin and race responses in administrative records compared to the 2010 Census. Hispanics are less likely to have missing Hispanic origin data but more likely to have missing race data in administrative records. Non-Hispanic Asian and NHPI individuals are more likely to have missing race and Hispanic origin data in administrative records. Younger individuals, renters, single parent households, individuals living in households with two or more people, individuals who responded to the census in the nonresponse follow-up operation, and individuals residing in the West are more likely to have non-matching race and Hispanic origin responses. Younger individuals, individuals living in households with two or more people, and nonresponse follow-up respondents are more likely to have missing race and Hispanic origin responses.

REFERENCES

Arday, Susan L., David R. Arday, Stephanie Monroe, and Jianyi Zhang. 2000. "HCFA’s Racial an Ethnic Data: Current Accuracy and Recent Improvements." *Health Care Financing Review* 21(4): 107-116.

Bentley, Michael, Tracy Mattingly, Christine Hough, and Claudette Bennett. 2003. “Census Quality Survey to Evaluate Responses to the Census 2000 Question on Race: An Introduction to the Data.” Census 2000 Evaluation B.3. Washington, DC: U.S. Census Bureau.

Bhaskar, Renuka, Adela Luque, Sonya Rastogi, and James Noon. "Coverage and Agreement of Administrative Records and 2010 American Community Survey Demographic Data." Forthcoming 2014.

Brown, J. Scott, Steven Hitlin, and Glen H. Elder, Jr. 2006. "The Greater Complexity of Lived Race: An Extension of Harris and Sim." *Social Science Quarterly* 87(2): 411-431.

Bye, Barry and Dean Judson. 2004. “Results From the Administrative Records Experiment in 2000.” Census 2000 Synthesis Report No. 16, U.S. Census Bureau.

Campbell, Mary E. and Christabel L. Rogalin. 2006. “Categorical Imperatives: The Interaction of Latino and Racial Identification.” *Social Science Quarterly* 87(5):1030-1052.

Compton, Elizabeth, Michael Bentley, Sharon Ennis and Sonya Rastogi. 2012. “2010 Census Race and Hispanic Origin Alternative Questionnaire Experiment.” DSSD 2010 CPEX Memorandum Series #B-05-R, 2010 Census Planning Memoranda Series #211, U.S. Washington, DC: U.S. Census Bureau.

del Pinal, Jorge and Dianne Schmidley. 2005. “Matched Race and Hispanic Origin Responses from Census 2000 and Current Population Survey February to May 2000.” Population Division Working Paper No. 79. Washington, DC: U.S. Census Bureau.

Doyle, Jamie M. and Grace Kao. 2007. “Are Racial Identities of Multiracials Stable? Changing Self-Identification among Single and Multiple Race Individuals.” *Social Psychology Quarterly* 70(4):405-423.

Dusch, Gianna and Fred Meier. 2012. “2010 Census Content Reinterview Survey Evaluation Report.” 2010 Census Program for Evaluations and Experiments. Washington, DC: U.S. Census Bureau.

Eschbach, Karl and Christina Gomez. 1998. "Choosing Hispanic Identity: Ethnic Identity Switching among Respondents to High School and Beyond." *Social Science Quarterly* 79(1): 74-90.

Farber, James, Deborah Wagner, and Dean Resnick. 2005. “Using Administrative Records for Imputation in the Decennial Census.” *Proceedings of the 2005 Joint Statistical Meetings,* Survey Research Methods Section, Alexandria, VA: American Statistical Association.

Farber, James and Charlene Leggieri. 2002. “Building and Validating a National Administrative Records Database for the United States.” New Zealand Conference on Database Integration.

Fernandez, Leticia, Sonya Rastogi, Sharon R. Ennis, and James Noon. 2015. " Evaluating Race and Hispanic Origin Responses of Medicaid Participants Using Census Data." Center for Administrative Records Research and Applications Working Paper #2015-XX, U.S. Census Bureau.

Gomez, Scarlett L., Jennifer L. Kelsey, Sally L. Glaser, Marion M. Lee, and Stephen Sidney. 2005. "Inconsistencies between Self-Reported Ethnicity and Ethnicity Recorded in a Health Maintenance Organization." *Ann Epidemiol* 15(1): 71-79.

Humes, Karen, Nicholas Jones, and Roberto Ramirez. 2011. "Overview of Race and Hispanic Origin: 2010." 2010 Census Brief C2010BR-02. Washington, DC: U.S. Census Bureau.

Jones, Nicholas A. and Jungmiwha Bullock. “The Two or More Races Population: 2010.” Census 2010 Brief C2010BR-13. Washington, DC: U.S. Census Bureau.

Kressin, Nancy R., Bei-Hung Chang, Ann Hendricks, and Lewis E. Kazis. 2003. "Agreement between Administrative Data and Patients’ Self-Reports of Race/Ethnicity." *American Journal of Public Health* 93(10): 1734-1739.

Liebler, Carolyn A., Sonya Rastogi, Leticia E. Fernandez, James Noon, and Sharon R. Ennis. 2014. "America’s Churning Races: Race and Ethnic Response Changes between Census 2000 and the 2010 Census." Center for Administrative Records Research and Applications Working Paper #2014-03, U.S. Census Bureau.

Lofquist, Daphne, Terry Lugaila, Martin O’Connell, and Sarah Feliz. “Households and Families: 2010.” Census 2010 Brief C2010BR-14. Washington, DC: U.S. Census Bureau.

McAlpine, Donna D., Timothy J. Beebe, Michael Davern, and Kathleen T. Call. 2007. "Agreement between Self-Reported and Administrative Race and Ethnicity Data among Medicaid Enrollees in Minnesota." *Health Services Research* 42(6): 2373-2388.

Obenski, S., Farber, J., and Chappell, G.. 2005. “Research to Improve Census Imputation Methods: Item Results and Conclusions.” *Proceedings of the 2005 Joint Statistical Meetings*, Survey Research Methods Section, Alexandria, VA: American Statistical Association.

Perez, Anthony Daniel. 2008. "Who is Hispanic? Shades of Ethnicity among Latino/a Youth." Pp. 17-35 in *Racism in Post-Race America: New Theories, New Directions*. Charles Gallagher (ed). Chapel Hill, NC: Social Forces.

Rastogi, Sonya, Leticia Fernandez, James Noon, Ellen Zapata, and Renuka Bhaskar. 2014. "Exploring Administrative Records Use for Race and Hispanic Origin Item Non-Response." Center for Administrative Records Research and Applications Working Paper #2014-16, U.S. Census Bureau.

Rastogi, Sonya and Amy O'Hara. 2012. "2010 Census Match Study." 2010 Census Planning Memoranda Series #247, U.S. Census Bureau.

Rothhaas, Cynthia, Frederic Lestina, and Joan M. Hill. 2012. "2010 Census Item Nonresponse and Imputation Assessment Report." 2010 Census Planning Memoranda Series #173, U.S. Washington, DC: U.S. Census Bureau.

Singer, Phyllis and Sharon R. Ennis. 2003. "Census 2000 Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview." Census 2000 Evaluation B.5.Washington, DC: U.S. Census Bureau.

U.S. Census Bureau; 2010 Census Summary File 1; Table P5; generated by Sharon Ennis; using American FactFinder; <http://factfinder2.census.gov>; (16 January 2015).

U.S. Census Bureau. 1993. "Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview." 1990 Census of Population and Housing Evaluation and Research Reports. Washington, DC: U.S. Census Bureau.

Wagner, Deborah and Mary Layne. 2012. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research & Applications Record Linkage Software." Center for Administrative Records Research and Applications Working Paper #2014-01, U.S. Census Bureau.

Zaslavsky, Alan M., John Z. Ayanian, and Lawrence B. Zaborski. 2012. The Validity of Race and Ethnicity in Enrollment Data for Medicare Beneficiaries. *Health Services Research* 47(3): 1300-1321.