Accuracy and Bias of Race/Ethnicity Codes in the Medicare Enrollment Database

Daniel R. Waldo, M.A.

Medicare administrative data are fairly accurate in identifying people who affiliate with White or Black racial groups; but less so for other race groups or for Hispanic/ Latino origin. Some differences were found between people who were identified as members of these other race groups and those who were missed by the administrative data. Although Medicare administrative files are a useful source of data for analysis of disparities in health care, researchers should be careful to use alternate data sources to test for potential differences between identified and unidentified members of racial and ethnic groups in the attributes being studied.

INTRODUCTION

Medicare's administrative files have the potential to be a rich source of information about disparities in health care access, use, or outcomes. Detailed information can be extracted on services used, diagnosis, treatment, and program expenditures. Used alone or in combination with other sources, the administrative files can be of enormous help in macro- and micro-analysis of disparities.

The inclusion of race and ethnicity in such analysis is almost universal, but not without problems. It has been argued, for example, that using race as a variable in social science research distracts one's attention from the life factors that are causally related to differences in health outcomes (Fullilove, 1998). Others have contended that race and ethnicity categories are too aggregated to be meaningful (Bhopal and Donaldson, 1998; Kaufman, 1999). These concerns notwithstanding, race and ethnicity endure among the most frequently used demographic characteristics in exploratory analysis of potential disparities in health care (e.g., Gornick, Eggers, and Reilly, 1996; Eggers and Greenberg, 2000). They act as much to proxy characteristics (such as culture, social networks, and socioeconomic status) that cannot be captured in administrative data, as to detect the role of race as an overt barrier to health care.

No matter how one uses race and ethnicity in analysis of differences in program experience, that analysis can easily be confounded if race and ethnicity are not well identified in the administrative records. This inquiry extends an earlier examination of the Medicare data (Arday et al., 2000) and has two goals: to review the accuracy of race and ethnicity classification in Medicare administrative data, and to help researchers identify potential problems when using those administrative data to examine disparities in health and health care. How accurate are Medicare data files as a source of information on race and ethnicity? Are there differences between those members of racial and ethnic groups who are identified as such in the administrative data, and those who are not identified?

The author is with the Centers for Medicare & Medicaid Services (CMS). The statements expressed in this article are those of the author and do not necessarily reflect the views or policies of CMS.

RACE/ETHNICITY IN MEDICARE PROGRAM DATA

At present, race and ethnicity are recorded in Medicare's administrative files with a single variable. This variable can take one of six different values—White, Black, Asian, North American Native, Hispanic, and other (plus an unknown category). Only one category is coded for each person in the file. This variable is populated principally with data provided to the Social Security Administration (SSA) by people at the time they apply for a Social Security number, when they apply or re-apply for benefits, or when they apply for a replacement Social Security card. Reporting race is voluntary and is not checked or verified by the Social Security staff receiving the application. Data are transferred from SSA's databases to the Medicare enrollment database (EDB), maintained at CMS.

Previous researchers have documented a number of problems with race coding in the EDB, concluding that these files were especially prone to difficulties in identifying people who affiliate with Hispanic/Latino ethnicity or with race groups other than White or Black. Comparing the EDB distribution of race/ethnicity with U.S. Bureau of the Census figures and with SSA data on country of birth, Lauderdale and Goldberg (1996) concluded that the EDB captures one-quarter of people with Hispanic heritage, less than one-fifth of American Indian enrollees, and just over one-half of Asian enrollees (with marked differences in success at capturing subracial groups in the Asian population). Pan and colleagues (1999) compared Medicare and Medicaid race codes for people in New Jersey enrolled in both programs, and found that agreement between the two administrative sets was highest for White and Black enrollees. Agreement was quite poor for Hispanic enrollees; too few Asian or American Indian enrollees were in the sample to draw conclusions about agreement on those codings. Arday and colleagues (2000) compared EDB race codes to enrollee self-reported race and ethnicity from the Medicare Current Beneficiary Survey (MCBS). They found that although agreement between the two sets of information in 1997 was an improvement over that in 1991, the degree of agreement for race groups other than White or Black remained low.

Part of the contribution of the work reported here is to assess whether there has continued to be a convergence of EDB race coding with that reported by enrollees themselves. More important, however, is its exploration for systematic differences in the characteristics and program experience of people missed by a racial code and those captured by that code—is it safe to use Medicare claims data to research racial disparities in health care?

DATA AND METHODS

Data

This study assesses the accuracy of and bias in EDB race codes by comparing them to the race and ethnicity affiliation self-reported by participants in the MCBS. The MCBS is a continuous survey of a nationally representative sample of Medicare enrollees. Survey respondents are interviewed three times each year for 3 years, plus an opening and closing interview. Information is collected from respondents about social, economic, and demographic life factors, about their use of health care services, and about financing of those services. This survey information is combined with administrative data on use of and payment for medical care through the Medicare Program (Adler, 1994).

Table 1

Self-Reported Race and Ethnicity Affiliations of Medicare Current Beneficiary Survey Participants: 1998-2001

Race	Hispanic/Latino Origin			
	Yes	No	Don't Know or Refused	Total
Native Hawaiian or Other Pacific Islander	37	179	0	216
Asian	10	315	0	325
American Indian or Alaskan Native	43	537	1	581
Black or African American	12	3,016	7	3,172
White	1,822	26,009	23	27,854
Other	289	38	2	329
No Affiliation (Replied Don't Know or Refused to All Categories)	9	15	13	37
Unduplicated Total	2,288	29,706	44	32,038

NOTES: Many of the people reporting "other" race specified particular countries of origin, especially Puerto Rico and countries from Central or South America. People reporting more than one racial affiliation were included in each group with which they affiliated; the unduplicated total includes each person only once.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Current Beneficiary Survey Access to Care Files, 1998-2001.

The categorization of race and ethnicity in the MCBS is different from that in the EDB. During the first interview, community-dwelling survey participants are asked whether they consider themselves to be of Hispanic or Latino origin. They are also asked to affiliate themselves with race categories; beginning in 1998, these categories conform to guidelines established by the Office of Management and Budget (1997): American Indian or Alaska Native, Asian, Black, Native Hawaiian or Other Pacific Islander, and White. Also beginning in 1998, respondents may identify more than one category, and may identify themselves as of "other" race as well. If a proxy is responding for a community-dwelling participant, the proxy is asked these questions; for participants in facilities, the facility staff person completing the interview is asked.

In this analysis, data from the 1998, 1999, 2000, and 2001 MCBS Access to Care Files (Centers for Medicare & Medicaid Services, 2002) were combined, to increase the number of participants affiliating with categories other than White or Black. Proxy responses and responses for people in facilities were dropped, resulting in a pool of 32,038 participants.[1] Only the most recent data were retained for those people who were included in more than one of the Access to Care Files.

For the purposes of this study, the MCBS was considered the authoritative source of race and ethnicity information on survey participants. This is because respondents were asked specifically to make their affiliations known, and because the MCBS permits a distinction between race and Hispanic ethnicity by asking separate questions. Table 1 shows how the study participants reported race and ethnicity.

Methods

To test the accuracy of EDB race coding, each of the six race codes was examined in turn. To the extent that the race/ethnicity code in the administrative data matches MCBS participants' self-identified affiliations, the former source can be said to be accurate.

For each code, MCBS participants were sorted into one of four groups (Arday et al., 2000):

- True positive (identified): The EDB assigns the MCBS participant to this category, and the participant affiliates with this category in the survey.

[1] However, of those participants, racial identification was not obtained for 37 respondents, and Hispanic origin was not obtained for 44 respondents.

- False positive (misidentified): The EDB assigns the MCBS participant to this category, but the participant does not mention this affiliation in the survey.
- False negative (missed): The EDB does not assign the MCBS participant to this category, but the participant affiliates with the category in the survey.
- True negative: The EDB does not assign the MCBS participant to this category, nor does the participant affiliate with the category in the survey.

People who affiliated with Native Hawaiian or Pacific Islander races in the MCBS were included with people who affiliated with the "other" race category. People who identified themselves as of Hispanic origin were compared to the EDB Hispanic race category regardless of the race affiliations they mentioned in the survey; in this way, ethnicity and race are collapsed into a single categorization, consistent with the concept used in the EDB.

Five measures of accuracy were calculated for each race code (Arday et al., 2000).

- Sensitivity (the ratio of true positives to true positives plus false negatives) reflects the likelihood that the EDB correctly identifies a person of the given race/ethnicity.
- Specificity (the ratio of true negatives to true negatives plus false positives) reflects the likelihood that it correctly excludes a person not of that race/ethnicity.
- Positive predictive value (the ratio of true positives to true positives plus false positives) reflects the likelihood that a person so coded in the EDB affiliates with that race/ethnicity.
- Negative predictive value (the ratio of true negatives to true negatives plus false negatives) reflects the likelihood that a person not so coded does not affiliate with that race/ethnicity.

The higher each of these ratios is, the more accurate the EDB can be said to be. In addition, a kappa value (Watkins and Pacheco, 2000) was computed for each of the EDB race categories to quantify the degree of agreement between the MCBS and EDB reports of race and ethnicity.

After examining the degree of agreement between the EDB and MCBS race codes, logistic regression analysis (using Stata™ version 6) was used to determine whether missed members of the group differ from the people identified with that race in the EDB. That is, is there a bias in the EDB race codes in various dimensions of the relevant populations? For each EDB code, true positive and false positive cases were combined to a category called "captured;" these people comprised the group against which members of the false negative (missed) group were compared (true negative respondents were omitted). A categorical variable was created with its value set to zero for the captured group and to one for the missed group and was used as the dependent variable. Age group, sex, metropolitan/non-metropolitan residence, income, and educational attainment were used as explanatory variables.

Several other regressions were performed as well. In these regressions, measures of health status and satisfaction with the quality of care received were used as dependent variables, as were program use measures for fee-for-service (FFS) enrollees. In addition to the categorical variable differentiating between captured and missed people, age group, sex, and metropolitan/non-metropolitan residence were used as explanatory variables. If the regression coefficient assigned to the categorical race variable were significantly different from zero, we could conclude that missed members of a racial or ethnic group differed from those so identified in the EDB beyond the other factors included in the regression. Table 2 outlines these regression analyses.

Table 2

Regression Models to Test Bias in Enrollment Database (EDB) Race Codes: 1998-2001

Area	Method	Dependent Variables	Independent Variables[1]
Race Capture	Logistic	Missed by EDB Race Code (Compared to Captured)	Age Strata Sex Metropolitan Residence Income Strata Educational Attainment General Health Status
Health Status	Logistic	Ever Told Survey Participant They Had Diabetes Ever Told Survey Participant They Had Hypertension, etc.	Age Strata Sex Metropolitan Residence Income Strata General Health Status
	Ordered Logistic	Self-Assessed Health Status	
System Satisfaction	Ordered Logistic	Satisfaction with Quality of Care Received Last Year	Age Strata Sex Metropolitan Residence Income Strata Educational Attainment General Health Status
Program Use	Logistic	In Managed Care In Medicaid Used Inpatient Services Used Outpatient Services Used Home Health Services Used Physician Services	Age Strata Sex Metropolitan Residence EDB Race Capture
	Ordinary Least Squares	(Natural Logarithms) Part A Reimbursement Part B Reimbursement	

[1] All independent variables are treated as categorical.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Current Beneficiary Survey Access to Care Files, 1998-2001.

No attempt was made to adjust variance estimates to reflect the complex stratified sample design of the MCBS. The categorical age variable used in the analysis mirrors the stratification of the sample design, removing that effect from the regression results. Pooling participants from three different years of the survey makes it very difficult to determine the appropriate correction factor for the effect of clustered sampling. In any event, the bias in variance estimates is likely to be in the downward direction, so findings that are statistically insignificant using uncorrected variance estimates would remain insignificant could an unbiased variance estimate be calculated. Findings that appear to be statistically significant using uncorrected variance estimates must be examined more closely.

RESULTS

Accuracy of EDB Race Codes

Comparing EDB codes with self-identified racial and ethnic affiliation confirms the weakness of the administrative data found by Lauderdale and Goldberg (1996) with respect to people who describe themselves as Hispanic or Latino, and for those who affiliate with American Indian or Asian races (Arday et al., 2000).

Table 3

Accuracy of Enrollment Database Race Codes Compared to Medicare Current Beneficiary Survey (MCBS) Codes: 1998-2001

Race	Enrollment Database	MCBS		Accuracy and Agreement Measures				
		Yes	No	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Kappa
						Percent		
White	Yes	26,498	481	96.5	88.2	98.2	79.1	0.81
	No	951	3,609					
Black	Yes	2,931	107	95.6	99.6	96.5	99.5	0.96
	No	134	28,367					
Asian	Yes	170	73	54.0	99.8	70.0	99.5	0.61
	No	145	31,151					
North American Native	Yes	41	18	20.6	99.9	69.5	99.5	0.32
	No	158	31,322					
Other	Yes	30	290	5.9	99.1	9.4	98.5	0.06
	No	481	30,738					
Hispanic	Yes	817	21	35.7	99.9	97.5	95.3	0.50
	No	1,471	29,685					

NOTES: Race comparisons only include those participants who reported a single racial affiliation; comparison of Hispanic/Latino included people with more than one racial affiliation. Analysis limited to those MCBS participants who reported directly. MCBS people who reported Native Hawaiian or Pacific Islander races were included with "other."

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Current Beneficiary Survey Access to Care Files, 1998-2001.

In this part of the study, only people who affiliated with a single race in the MCBS were analyzed. This included the vast majority of participants: Overall, only 1 percent of respondents reported more than one race affiliation (not counting Hispanic/Latino origin). People who affiliated with American Indian were far more likely to consider themselves multiracial than were people who did not do so—two-thirds of them affiliated with a second race (mostly White).

The comparison of EDB and MCBS race codes is summarized in Table 3 for people who affiliated with only one race code in the MCBS. Sensitivity levels were high for White and Black EDB codes, indicating that a large proportion of those enrollees were correctly captured in the EDB, but sensitivity levels were quite low for other groups. Only one-half of people who identified themselves as Asian were correctly identified in the EDB, only one-third of those who thought of themselves as Hispanic/Latino, and only one-fifth of those who reported being American Indian. The positive predictive value of the White, Black, and Hispanic EDB codes was high (that is, the EDB was accurate in those it placed in the groups); that for Asian and North American Native was considerably lower, dropping to 64 percent for the latter (one out of every three people identified as North American Natives in the EDB did not affiliate that way in the MCBS). Standard interpretation of kappa values (Watkins and Pacheco, 2000) indicates strong agreement of the EDB and MCBS on the categories of White and Black, fair agreement on the categories of Asian and Hispanic, and poor agreement on the category of American Indian.

Sociodemographic variables appear to have little predictive power when it comes to EDB identification of race and ethnicity.

Logistic regression of the race identification variable (missed people compared to captured people) on age, sex, urban/rural residence, income, education, and general health accounts for between 8 and 12 percent of the total variation in the dependent variable (Table 4). People who affiliate with America Indian and live in rural areas were one-half as likely to be unidentified as were their urban counterparts. However, when attention is restricted only to those who affiliated with American Indian and no other race, that effect becomes statistically insignificant, and the effect of sex and of educational attainment becomes statistically significant: females were one-third as likely to be missed as males, and those with more than a high school diploma were one-third as likely to be missed as those with a high school diploma only. People who affiliated with Asian race and whose income was $25,000 or more were twice as likely to be missed as those with income between $10,000 and $25,000. The odds of people being missed who identified themselves as of Hispanic origin increases with the age of the person; high-income Hispanic people were more likely to be missed, as were those with less than a high school education.

Bias in the EDB Race Code

Given that the EDB misses a substantial proportion of people who affiliate with American Indian and Asian race categories and with Hispanic/Latino heritage, it is important to determine the extent to which those it does categorize in these groups are similar to those it has missed. If the groups are similar, then the EDB can be used to compare the average experience of enrollees of different racial and ethnic groups, even if it cannot illuminate total levels of expenditure for the groups.

Regression results testing this question are summarized in Tables 5, 6 and 7. In this part of the study, people were included in the analysis even if they affiliated with more than one race; if a participant affiliated with both American Indian and Asian races, for example, the participant was included in both regressions. All people who affiliated with Hispanic/Latino origin were included in that regression analysis. Explanatory variables were limited to the sociodemographic characteristics available from the EDB: age, sex, and rural/urban residence. This allowed simulation of the effects one would encounter when using only the EDB as a source of information about the beneficiaries being studied.

Differences between captured and missed survey participants varied among the EDB race codes in aspects of their health (Tables 5 and 6). American Indians missed by the EDB were five times as likely to have been diagnosed with skin cancer as people captured by that race code. Missed Asian people were one-quarter as likely as to have had angina or coronary heart disease as the people identified as Asian in the EDB. People of Hispanic heritage missed by the EDB were less likely to have been diagnosed with rheumatoid arthritis, and more likely to been diagnosed with skin cancer or osteoporosis. They also tended to report better health status than those captured in the EDB as Hispanic.

People who affiliated with the Asian race or with Hispanic heritage, but who were missed by those EDB codes, exhibited some differences from the people captured by those codes in terms of Medicare Program interactions (Tables 6 and 7). Missed Asian survey participants were twice as likely to be in a managed care plan, one-quarter as likely to be in Medicaid, and (if in FFS Medicare) one-

Table 4

Results of Logistic Regression Comparing People Missed by an Enrollment Database (EDB) Race Code to Those Captured by the Code, on Selected Sociodemographic Characteristics: 1998-2001

Independent Variable	American Indian		Asian	Hispanic
	Single/Multiple Race Affiliation	Single Race Affiliation	Single/Multiple Race Affiliation	Single/Multiple Race Affiliation
	Odds Ratio			
Age[1]				
45-64 Years	NS	NS	NS	1.68 (p=0.019)
65-69 Years	NS	NS	NS	4.26 (p<0.001)
70-74 Years	NS	NS	NS	NS
75-79 Years	NS	NS	NS	NS
80-84 Years	NS	NS	NS	NS
85 Years or Over	NS	NS	NS	2.85 (p<0.001)
Sex[2]				
Female	NS	0.36 (p=0.011)	NS	NS
Place of Residence[3]				
Non-Metropolitan	0.44 (p=0.010)	NS	NS	NS
Income[4]				
Unknown	NS	NS	NS	NS
> $25,000	NS	NS	2.29 (p=0.003)	2.25 (p<0.001)
Educational Attainment[5]				
Unknown				
< High School	NS	NS	NS	0.71 (p=0.010)
> High School	NS	0.31 (p=0.026)	NS	NS
Self-Assessed Health Status [6]				
Unknown	NS	NS	NS	NS
Poor/Fair	NS	NS	NS	NS
Very Good/Excellent	NS	NS	NS	NS
Pseudo R^2	0.08	0.12	0.10	0.06
Number of Missed Persons[7]	424	126	118	1,051
Number of Captured Persons[8]	51	49	188	606

[1] Compared to under 45 years.

[2] Compared to male.

[3] Compared to metropolitan.

[4] Compared to < $25,000.

[5] Compared to high school graduate.

[6] Compared to good.

[7] Dependent variable is 1: the EDB code is not set, but the participant affiliates with that race/ethnicity in the Medicare Current Beneficiary Survey (MCBS).

[8] Dependent variable is 0: race/ethnicity under consideration is set in the EDB.

NOTES: Results are shown for the subset of MCBS participants who affiliated solely with American Indian race, as well as those who affiliated with American Indian and another race(s). "NS" means the odds ratio is not significantly different from 1.0 (p>0.05). For other cells, p value is shown in parentheses for clustered sampling effect.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Current Beneficiary Survey Access to Care Files, 1998-2001.

half as likely to have used Medicare physician services. Missed Hispanic survey participants were one-half as likely to have been in Medicaid. None of the regressions showed significant differences in total per capita Medicare Part A or Part B reimbursement between missed and captured populations.

Table 5

Odds of People Missed by an Enrollment Database Race Code Reporting Selected Health Conditions, Compared to Those Captured by the Race Code, Controlling for Selected Sociodemographic Characteristics: 1998-2001

Dependent Variable	Racial or Ethnic Category Considered		
	American Indian	Asian	Hispanic
	Odds Ratio		
Has a Doctor (Ever) Told You That You Had			
Diabetes, High Blood Sugar, Sugar in Your Urine?	NS	NS	NS
Emphysema, Asthma, or COPD?	NS	NS	NS
Rheumatoid Arthritis?	NS	NS	0.73 (p=0.017)
Arthritis (In Any Other Part of Your Body) Other Than Rheumatoid Arthritis?	NS	NS	NS
Skin Cancer?	5.12 (p=0.026)	NS	1.70 (p=0.008)
Any (Other) Kind of Cancer, Malignancy, or Tumor Other Than Skin Cancer?	NS	NS	NS
Angina Pectoris or Coronary Heart Disease?	NS	0.23 (p=0.008)	NS
A Myocardial Infarction or Heart Attack?	NS	NS	NS
Hypertension, Sometimes Called High Blood Pressure?	NS	NS	NS
A Stroke, a Brain Hemorrhage, or a Cerebro-Vascular Accident?	NS	NS	NS
Osteoporosis, Sometimes Called Fragile or Soft Bones?	NS	NS	1.36 (p=0.040)
A Mental or Psychiatric Disorder, Including Depression?	NS	NS	NS

NOTES: "NS" means the odds ratio is not significantly different from 1.0 (p>0.05). For other cells, p value is shown in parentheses, not adjusted for clustered sampling effect. Independent variables not shown comprise age stratum, sex, and urban/rural residence. COPD is chronic obstructive pulmonary disease.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Current Beneficiary Survey Access to Care Files, 1998-2001.

Table 6

Results of Multivariate Regression of Selected Beneficiary Measures on Selected Sociodemographic Characteristics, Showing the Effect of Having Been Missed by an Enrollment Database (EDB) Race Code: 1998-2001

Dependent Variable	Racial or Ethnic Category Considered		
	American Indian	Asian	Hispanic
	Multinomial Regression Coefficient		
In general, compared to other people your age, would you say that your health is excellent, very good, good, fair, or poor?	NS	NS	0.2912 (p=0.001)
Please tell me how satisfied you have been with the overall quality of the medical services you have received over the past year (very satisfied, satisfied, dissatisfied, very dissatisfied).	NS	NS	NS
	Ordinary Least Squares Regression Coefficient		
Logarithm of Medicare Part A Reimbursement (Fee-for-Service Enrollees Only)	NS	NS	NS
Logarithm of Medicare Part B reimbursement (Fee-for-Service Enrollees Only)	NS	NS	NS

NOTES: "NS" means the coefficient is not significantly different from zero (p>0.05). For other cells, p value is shown in parentheses, not adjusted for clustered sampling effect. Independent variables not shown comprise age stratum, sex, and urban/rural residence.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Current Beneficiary Survey Access to Care Files, 1998-2001.

DISCUSSION

Generally, the measures of accuracy reported in this study are better than those calculated using the 1996 MCBS data (Arday et al., 2000). This is partly because of improvements to the EDB previously discussed, and may be due in part to changes starting in 1998 in the way the MCBS race and ethnicity questions were asked. However, while the specificity of the EDB codes (relatively low false positives) is generally good, their sensitivity (relatively low false negatives) for categories

Table 7

Odds of People Missed by an Enrollment Database Race Code Exhibiting Selected Medicare Program Characteristics, Compared to Those Captured by the Race Code, Controlling for Selected Sociodemographic Characteristics: 1998-2001

Dependent Variable	Racial or Ethnic Category Considered		
	American Indian	Asian	Hispanic
		Odds Ratio	
Survey Participant was in a Managed Care Plan	NS	2.08 (p=0.001)	NS
Survey Participant was in Medicaid	NS	0.27 (p<0.001)	0.46 (p<0.001)
Fee-for-Service Survey Participant Used Inpatient Services	NS	NS	NS
Fee-for-Service Survey Participant Used Outpatient Services	NS	NS	NS
Fee-for-Service Survey Participant Used Home Health Services	NS	NS	NS
Fee-for-Service Survey Participant Used Physician Services	NS	0.46 (p=0.045)	NS

NOTES: "NS" means the odds ratio is not significantly different from 1.0 (p>0.05). For other cells, p value is shown in parentheses, not adjusted for clustered sampling effect. Independent variables not shown comprise age stratum, sex, and urban/rural residence.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Current Beneficiary Survey Access to Care Files, 1998-2001.

other than White and Black remain poor. Further, if people who reported more than one racial affiliation are included in the analysis, the sensitivity measure for American Indians drops markedly—only 5.4 percent of people who affiliated with American Indian in the MCBS are coded as such in the EDB.

It should be noted that high rates for specificity and negative predictive value reflect more the mechanical effect of the overall distribution of the race attributes in the Medicare population than the accuracy of the EDB. The relatively low specificity of the White category is consistent with a construction of that category as the comparison group (Bhopal and Donaldson, 1998).

There are two limitations to the MCBS here, especially regarding American Indians and Native Hawaiians. First, people who reported affiliation with more than one racial group were not asked about the relative strength of those affiliations: do they think of themselves principally as one race first and then as the other—and if so, which race is primary? Second, the MCBS geographic sampling areas include neither Alaska, Hawaii nor mainland Tribal reservations. Thus, it is likely that American Indians and Native Hawaiians are underrepresented in the MCBS sample. Further, it may very well be that those who were not in the MCBS sample frame have a very different likelihood of being correctly identified by race in the EDB than those who were in the sample frame. This supposition can be tested as results of administrative data matches with the Indian Health Service files are analyzed.

The findings support the hypothesis that unidentified members of a racial or ethnic group differ from those who are identified in the EDB, especially for Hispanic enrollees. The differences between missed members of these groups, and enrollees identified as being of the group appear in different aspects of program experience for different race or ethnicity categories, suggesting that before undertaking analysis of Medicare administrative data, researchers should test for the presence of a bias in dependent or independent variables, using alternate data sets such as the MCBS.

The MCBS does not contain information on the date when the EDB race code was established for the survey participant, leaving a potential gap in its ability to identify bias. Despite the improvements in race coding described, Lauderdale and Goldberg's (1996) note remains valid: recent immigrants or enrollees are more likely to have minority EDB codes than would their older counterparts. This is because the latter are more likely to have codes from the early

days of the program when race and ethnicity were less well documented. Lauderdale and Goldberg also argue that this introduces an immediate bias, as waves of immigrants tend to have different histories and health behaviors. The extent of this problem is not addressed here, and should be borne in mind when using the data as they exist.

CONCLUSIONS

Comparison of EDB race codes with the self-affiliations of MCBS participants reveals differing levels of accuracy and bias in Medicare administrative data. Generally speaking, the codes for White and Black enrollees are fairly accurate, but the codes for Asian and American Indian enrollees are much less accurate. The Hispanic race code captures only one-third of those who identify themselves as being of Hispanic/Latino origin. Further, there are statistically significant differences between those who are labeled with a race code in the EDB and those who are missed by that code, in a number of aspects of health care use and insurance.

Initiatives to improve the accuracy of the race code may mitigate some of those problems. SSA data are now used annually to update the EDB, and CMS, and the Indian Health Service are exchanging administrative data to identify Medicare enrollees who are members of federally recognized Tribes. Work is currently underway to compare self-identified race from program surveys such as the Medicare Health Outcomes Survey and the Consumer Assessment of Health Plan Survey, with each other and with the EDB; based on the outcome of the comparison, protocols may be developed to change the enrollees' code in the EDB.

Medicare data remains a rich source of information for research into disparities in health care, but researchers should be cautious when using the EDB race code to conduct such studies. The problems posed by inaccurate or biased coding are not necessarily fatal, but could become so if left undiagnosed; fortunately, diagnostic tools such as the MCBS exist. Work in progress could improve coding in EDB, but this work will result in an evolution in the race codes—not a revolution.

ACKNOWLEDGMENTS

The author would like to thank Rosemarie Hakim, and colleagues in the Office of Research, Development, and Information for their helpful comments and suggestions on an earlier draft.

REFERENCES

Adler, G.S.: A Profile of the Medicare Current Beneficiary Survey. *Health Care Financing Review* 15(4):153-163, Summer 1994.

Arday, S.L., Arday, D.R., Monroe, S., et al.: HCFA's Racial and Ethnic Data: Current Accuracy and Recent Improvements. *Health Care Financing Review* 21(4):107-116, Summer 2000.

Bhopal, R. and Donaldson, L.: White, European, Western, Caucasian, or What? Inappropriate Labeling in Research on Race, Ethnicity, and Health. *American Journal of Public Health* 88(9):1303-1307, September 1998.

Centers for Medicare & Medicaid Services: The Medicare Current Beneficiary Survey, 2002. Internet address: www.cms.hhs.gov/mcbs. (Accessed July 2003).

Eggers, P.W. and Greenberg, L.G.: Racial and Ethnic Differences in Hospitalization Rates Among Aged Medicare Beneficiaries. *Health Care Financing Review* 21(4):81-95, Summer 2000.

Fullilove, M.T.: Abandoning Race as a Variable in Public Health Research: An Idea Whose Time Has Come. *American Journal of Public Health* 88(9):297-1298, September 1998.

Gornick, M.E., Eggers, P.W., and Reilly, T.W.: Effects of Race and Income on Mortality and Use of Services Among Medicare Beneficiaries. *New England Journal of Medicine* 335(11):791-799, September 12, 1996.

Kaufman, J.S.: How Inconsistencies in Racial Classification Demystify the Race Construct in Public Health Statistics. *Epidemiology* 10(2):101-103, March 1999.

Lauderdale, D.S. and Goldberg, J.: The Expanded Racial and Ethnic Codes in the Medicare Data Files: Their Completeness of Coverage and Accuracy. *American Journal of Public Health* 86(5):712-716, May 1996.

Office of Management and Budget: Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, 1997. Internet address: www.whitehouse.gov/omb/fedreg/ombdir15.html. (Accessed July 2003).

Pan, C.X., Glynn, R.J., and Mogun, H.: Definition of Race and Ethnicity in Older People in Medicare and Medicaid. *Journal of the American Geriatrics Society* 47(6):730-733, June 1999.

Watkins, M.W. and Pacheco, M.: Interobserver Agreement in Behavioral Research: Importance and Calculation. *Journal of Behavioral Education* 10(4):205-212, December 2000.

Reprint Requests: Daniel R. Waldo, M.A., Centers for Medicare & Medicaid Services, 7500 Security Boulevard, Mail Stop: C3-16-27, Baltimore, MD 21244-1850. E-mail: dwaldo@cms.hhs.gov