



Implementing the HEDIS<sup>®</sup>  
Medicare Health Outcomes Survey

**Imputation Analysis for HOS Income Data**

**Phase One: External Imputation**

Prepared by:

Joachim Bruess, PhD, Director of Analysis  
Myriam Bikah, Health Care Analyst

Prepared by the National Committee for Quality Assurance (NCQA)

for the Centers for Medicare and Medicaid Services (CMS)

Under Contract Number HHSM-500-2004-00015I-0001

OY2 Task 4.18b – Deliverable 411

October 13, 2008

Implementing the HEDIS<sup>®</sup>  
Medicare Health Outcomes Survey  
Imputation Analysis for HOS Income Data

**TABLE OF CONTENTS**

---

1.0	Background and Purpose.....	1
2.0	Imputing Missing Income Information .....	1
3.0	Income Imputation Methodology .....	3
4.0	Results.....	6
5.0	Conclusion.....	15
6.0	Figures.....	16

## **1.0 BACKGROUND AND PURPOSE**

Data from the Medicare Health Outcomes Survey (HOS) offer important longitudinal information about and insights into the self-assessed health status of the older population in the United States. CMS has surveyed this population annually over several years and repeatedly found that income information is missing in Baseline and Follow-up surveys for about 10 to 20 percent of respondents. To address this issue and to generate complete information across HOS data sets, CMS is collaborating with NCQA to explore and help define a valid income imputation method using HOS 2000 Cohort 3 Baseline and HOS 2006 Cohort 9 Baseline data that could be applied to other HOS cohorts.

## **2.0 IMPUTING MISSING INCOME INFORMATION**

Various income imputation methods exist; some use very basic imputation algorithms and others base imputation on more complex considerations. On the whole, however, income imputation methods that use a constant, such as the average or median of a sample or a given population, to replace missing information can reduce income variation considerably. Provided that a small proportion of income data is missing (e.g., below 5 percent), the effect of imputation might be negligible. However, if 20 percent or more income information is missing, imputation reduces income variation substantially because potential variation is locked in an imputed constant. Therefore, it is important that the imputation method includes and helps maintain variation in the resulting data set.

There are two basic imputation methods: internal and external. *Internal imputation* uses existing information from the same data set. Specific algorithms are employed to impute missing information based on: a) income variables; b) related variables; (e.g., education); or c) unrelated variables (e.g., general health status) from the same data set. *External imputation* uses income information from different data sources, which in turn must be matched to the existing data set and prepared for imputation. Income information from Census data (<http://www.census.gov/>) and data released from the Bureau for Labor Statistics (<http://www.bls.gov/>) or the IRS (<http://www.irs.gov/>) provide public sources for income imputation. CMS and NCQA agreed to explore the external income imputation method based on publicly available income information from Census 2000 data for the HOS 2000 Cohort 3 Baseline data to match the year of the Census 2000 and the HOS 2006 Cohort 9 Baseline data (which offers recent data).

Actual imputation of income information can use a variety of matching variables. For this study, we matched Census 2000 household income information for the 65-years-and-older population with HOS 2000 Cohort 3 Baseline and HOS 2006 Cohort 9 Baseline data on zip code and state levels. Ideally, we wanted to differentiate for gender, education and ethnic group, but this information is usually based on head-of-household information. Furthermore, the factual income information might differ; for example, the external source might provide annual income information in dollars, whereas HOS income information is grouped into income brackets (e.g., \$20,000—\$29,999 annually).

### **3.0 INCOME IMPUTATION METHODOLOGY**

Census 2000 data are publicly available and contain median and aggregated household income, median and aggregated household income by age of householder, aggregated income for the population 15+ years of age, and median and aggregated earnings by gender for the population 16+ years of age. Income data are available from every person who responded to the Census, whether employed or not. Data are broken down by zip code, county, state, metropolitan area and other criteria. Average income is available for all persons living in each zip code, county, or state. Data are not available for every individual at the zip code level because of privacy issues. Group distinctions can be made between gender and age groups (e.g., 65+ population). For income imputation, Census 2000 data offer categorical differences that can be matched reasonably with HOS 2000 Cohort 3 Baseline and HOS 2006 Cohort 9 Baseline data (e.g., zip code, state, age group).

Because Census 2000 information is dated, we used HOS data from the same year, i.e., 2000 Cohort 3 Baseline data, as the baseline model. Imputation for HOS 2006 Cohort 9 Baseline data included salary inflation to adjust for changes over time. As a caveat, Census income data and HOS income data do not necessarily overlap. Income from HOS can be higher or lower compared to Census income data. In addition, aggregated on the state level, income data might differ in some states and overlap in others and these differences aggregated at the state level will be reported.

HOS 2006 Cohort 9 Baseline data offer information for the year 2006; thus, inflation must be considered when imputing income from Census 2000. According to a Congressional Research Service report for Congress published in 2006, an annual increase in income (regular earnings or Social Security benefits) of about 2 percent is realistic for the 65+ population.<sup>1</sup> In addition, we reviewed income information for the general population at the state-level from the Bureau of Labor Statistics ( [http://www.bls.gov/oes/oes\\_dl.htm](http://www.bls.gov/oes/oes_dl.htm) ). State-level median and average household income per occupation is available and income data for “All Occupations” were used. Based on these data, the increase in income is slightly higher and ranges between 3 percent and 5 percent across states. Hence, a third model using a 3 percent income increase was tested in addition to the baseline 2000 imputation and the Cohort 9 model, with a 2 percent annual increase from 2000—2006.

Our approach for income imputation used the Census 2000 median household income information for the 65+ population on the zip code and state levels.<sup>2</sup> Since Census 2000 data contains median income data in dollar amounts, income for the zip code and state level was re-coded into the same income brackets listed in the HOS data sets<sup>3</sup>. For each missing piece of

---

<sup>1</sup> See Patrick Purcell and Debra Whitman (2006) *Topics in Aging. Income of Americans Age 65 and Older, 1969 to 2004*, p.16 ff.

<sup>2</sup> Including state information is important because some zip codes reach across state borders. This is done to ensure efficient postal delivery and examples can be found in military facilities that span multiple states, or in remote areas where two areas of a state can be serviced in conjunction.

<sup>3</sup> HOS income brackets, higher income corresponds with higher codes: (1) less than \$5,000; (2) \$5,000—\$9,999; (3) \$10,000—\$19,999; (4) \$20,000—\$29,999; (5) \$30,000—\$39,999; (6) \$40,000—\$49,999; (7) \$50,000—\$79,999; (8) \$80,000—\$99,999; (9) \$100,000 or more per year.

income information, the re-coded median household income from the matching Census 2000 zip code and state was imputed. Imputation might not be entirely complete because Census and HOS data differ on the state level. For example, Census data do not include Puerto Rico and HOS does include Puerto Rico data.

## 4.0 RESULTS

Three income imputation models were analyzed:

1. A *baseline model* that contains imputed income information from Census 2000 into the HOS 2000 Cohort 3 Baseline data set.
2. A *current model* that contains imputed and *inflation-adjusted (2 percent per year)* income information from Census 2000 into the HOS 2006 Cohort 9 Baseline data set.
3. A *second model* that contains imputed and *inflation-adjusted (3 percent per year)* income information from Census 2000 into the HOS 2006 Cohort 9 Baseline data set.

To facilitate the presentation, results are sometimes reported on the state level to keep the amount of presented information manageable.

### 4.1 The Baseline Model

A total of 195,618 people over the age of 65 completed 80 percent of the HOS 2000 Cohort 3 Baseline survey, and on average, 11.6 percent of respondents did not report income. Across the states, missing information ranged from 6.7 percent in Puerto Rico to 16 percent in North Carolina. There was an outlier with 100 percent missing income information in Montana; the sole respondent did not disclose income information.

Matching income based on the zip code and state levels led to an imputation for 22,563 respondents. After imputation, only 814 respondents (0.4 percent) in the HOS 2000 Cohort 3 data set were missing income information. The income distribution for HOS 2000 Cohort 3

shows (Table 1) that 80 percent of imputed income is concentrated in income brackets between \$20,000 and \$39,999, whereas income of respondents who reported their income ranges primarily between \$5,000 and \$39,999 (68%). Thus, the overall income in HOS 2000 Cohort 3 will be slightly higher after imputation.

<b>Income Brackets</b>	<b>Reported Income (n=173,055)</b>	<b>Imputed Income (n=21,749)</b>
Less than \$5,000	3.86	0.03
\$5,000 - \$9,999	10.67	0.09
\$10,000 – \$19,999	27.29	8.72
\$20,000 – \$29,999	18.99	50.16
\$30,000 – \$39,999	10.91	30.36
\$40,000 – \$49,999	6.25	7.55
\$50,000 – \$79,999	6.57	2.92
\$80,000 - \$99,999	1.50	0.08
\$100,000 and more	1.77	0.08
Don't know	12.19	0

Aggregated results on the state level show that in most states imputation does not lead to changes in income levels (see Diagrams 1—5, Figures). After imputation, income increases in Arkansas and the District of Columbia (one income bracket higher), and in South Carolina, income decreases after imputation.

Comparison of HOS 2000 Cohort 3 and Census 2000 income data shows that income is the same in 28 of 50 states (see Diagrams 1—5, Figures). Census 2000 income levels are higher before and after imputation in Alaska, Alabama, Connecticut, the District of Columbia, Hawaii, Louisiana, Maryland, North Dakota, New Jersey, Nevada, Utah, Vermont, Washington and Wyoming. In contrast, HOS 2000 Cohort 3 Baseline income levels are higher before and after

imputation compared to Census 2000 income information in Mississippi, New Hampshire, South Carolina and South Dakota. These differences can affect results before and after imputation, depending on the state-specific combination of HOS and Census income information. For example, Census income was higher than HOS for Arkansas before imputation; here imputation led to increased HOS income levels.

The impact of the income variable was tested with and without imputed income in a multivariate regression model. The model attempts to explain whether, and to what extent, people's self-assessed health depends on education, gender, age and income, compared to their peers.<sup>4</sup>

Three models explore effects of income imputation for groups of respondents (Table 2). One model focuses on respondents who submitted income information; another model is developed only for respondents for whom income was imputed; and a final model shows results involving all respondents after imputation.

---

<sup>4</sup>Coding the variable ranges from (1) for excellent health to (5) for poor health. The higher the value, the worse off you feel you are, compared with your peers.

	<b>Respondents who reported income (n=170,519 <sup>a)</sup>)</b>	<b>Respondents with imputed income (n=19,236)</b>	<b>All Respondents after imputation (n=189,755)</b>
Intercept	2.587 <sup>b)</sup> (0.031 <sup>c)</sup> )	2.275 (0.097)	2.544 (0.030)
Age	0.011 (0.000)	0.015 (0.001)	0.011 (0.000)
Gender	0.023 (0.005)	-0.008 (0.015)	0.018 (0.005)
Education	-0.166 (0.002)	-0.157 (0.006)	-0.166 (0.002)
Income	-0.031 (0.001)	-0.059 (0.009)	-0.031 (0.001)
R-Square	0.066	0.057	0.065
a) Listwise deletion leads to smaller sample sizes in multiple regression models b) Parameter estimates (standard errors) c) Estimates shown in grey are not significant ( $p < 0.01$ )			

Those who reported having less education, a lower income, being older in age and female are more likely to perceive their health to be worse than that of their peers. A comparison shows that the effects of the income variable are identical in the first and the last model. The model for respondents for whom income was imputed shows similar coefficients. The only difference is that the gender effect is not significant. On the whole, results from the baseline analysis — Census 2000 income data used for imputation into HOS Cohort 3 Baseline data— suggest that the income imputation based on zip code and state matching does not distort results.

#### **4.2 Models that Include Inflation**

A total of 107,341 people over the age of 65 completed 80 percent of the HOS 2006 Cohort 9 Baseline survey; on average, 10.2 percent of respondents did not report income. Matching income based on the zip code and state levels led to imputation for 10,791 respondents. After imputation, only 197 respondents (0.2 percent) in the HOS 2006 Cohort 9 data had missing

income information. Compared across states, in the 2 percent inflation model missing information ranged from 5.3 percent in Louisiana to 16.4 percent in the District of Columbia. Outliers due to small sample size were found in Montana (20 percent), South Dakota (33.3 percent) and South Carolina (41.7 percent) (see Diagrams 6—10, Figures).

Aggregated state level results show that in 42 of 50 states, inflation-adjusted income imputation does not lead to changes in income levels (see Diagrams 6—10, Figures). In six states, which include: Hawaii; Michigan; Minnesota; North Dakota; Puerto Rico; and Washington, as well as the District of Columbia, income decreases by one income bracket after imputation compared to income before imputation. In South Carolina and South Dakota, income decreases by two income brackets. Small sample sizes and lower Census income compared to HOS income contribute to differences in some states.

A comparison of HOS 2006 Cohort 9 Baseline and 2 percent inflation-adjusted Census 2000 income data shows that income is the same in 19 of 50 states (see Diagrams 6—10, Figures) and inflation-adjusted Census income levels are higher in 23 of 50 states. Although, the higher levels did not systematically increase HOS income levels after imputation.

Aggregated state level results show for the 3 percent inflation model that in 42 of 50 states, inflation-adjusted income imputation does not lead to changes in income levels (see Diagrams 11—15, Figures). In six states, which include: Hawaii; Michigan; Minnesota; North Dakota; Puerto Rico; Washington; as well as the District of Columbia, income decreases by one income

bracket after imputation compared to income before imputation, and decreases by two income brackets in South Carolina and South Dakota. Small sample sizes and lower Census income compared to HOS income contribute to differences in some states.

A comparison of HOS 2006 Cohort 9 Baseline and 3 percent inflation-adjusted Census 2000 income data shows that income is the same in 12 of 50 states (see Diagrams 11—15, Figures) and inflation-adjusted Census income levels are higher in 29 of 50 states. Although, the higher levels did not systematically increase HOS income levels after imputation. Comparison of income levels between the two models shows, as expected, that income levels are higher for the 3 percent inflation-adjusted model, although this is only the case in nine states. Recoding income into the HOS income brackets seems to have captured most of the differences between the 2 percent and 3 percent models.

The income distribution for HOS 2006 Cohort 9 before imputation shows (Table 3) that 58 percent of reported income ranged between \$10,000 and \$39,999. The income distribution for respondents who did not submit income information is concentrated between \$20,000 and \$49,999 (90%) when income was inflation adjusted for 2 percent, and income ranged primarily between \$20,000 and \$49,999 (88%) when income was adjusted for 3 percent inflation per year.

<b>Income Brackets</b>	<b>Reported Income (n=96,353)</b>	<b>Imputed Income 2% adjusted (n=10,791)</b>	<b>Imputed Income 3% adjusted (n=10,791)</b>
Less than \$5,000	3.59	0.02	0.02
\$5,000 - \$9,999	8.07	0.17	0.09
\$10,000 – \$19,999	26.10	4.84	3.95
\$20,000 – \$29,999	19.70	38.58	29.92
\$30,000 – \$39,999	11.79	37.94	42.17
\$40,000 – \$49,999	7.26	13.18	15.97
\$50,000 – \$79,999	7.73	5.10	7.57
\$80,000 - \$99,999	1.92	0.17	0.26
\$100,000 and more	2.31	0.02	0.05
Don't know	11.52	0.00	0.00

The impact of the income variable was tested with and without imputed income in a multivariate regression model. The model attempts to explain whether, and to what extent, people's self-assessed health depends on education, gender, age, income and body mass index (BMI), compared to their peers (Table 4).

	<b>Respondents who reported income (n=92,017 <sup>a)</sup>)</b>	<b>Respondents with imputed income, 2% inflation adjusted (n=9,708)</b>	<b>Respondents with imputed income, 3% inflation adjusted (n=9,708)</b>	<b>All Respondents after imputation, 2% inflation adjusted (n=101,725)</b>	<b>All Respondents after imputation, 3% inflation adjusted (n=101,725)</b>
Intercept	1.664 <sup>b)</sup> (0.048)	0.880 (0.155)	0.892 (0.155)	1.583 (0.046)	1.586 (0.046)
Age	0.014 (0.000)	0.022 (0.002)	0.022 (0.002)	0.015 (0.000)	0.015 (0.000)
Gender	0.014 (0.007) <sup>c)</sup>	0.021 (0.020)	0.022 (0.020)	0.013 (0.006)	0.013 (0.006)
Education	-0.158 (0.003)	-0.144 (0.008)	-0.143 (0.008)	-0.158 (0.002)	-0.157 (0.002)
Income	-0.038 (0.001)	-0.062 (0.011)	-0.063 (0.011)	-0.038 (0.001)	-0.038 (0.001)
BMI	0.025 (0.001)	0.030 (0.002)	0.030 (0.002)	0.025 (0.001)	0.025 (0.001)
R-Square	0.086	0.087	0.087	0.086	0.086
a) Listwise deletion leads to smaller sample sizes in multiple regression models b) Parameter estimates (standard errors) c) Estimates shown in grey are not significant (p<0.01)					

Those who reported having less education, lower income, being older in age and having a higher BMI are more likely to perceive their health to be worse than that of their peers. Comparison shows that the effects of the income variable are identical in the first and final models; they are substantially higher in the models for respondents for whom income was imputed.

The gender effect is not significant in all regression models. On the whole, results from this model, Census 2000 data inflation-adjusted by 2 percent, suggest that income imputation based on zip code and state matching does not distort results. Differences between the 2 percent and the 3 percent inflation-adjusted models are negligible and do not justify a strong recommendation one way or the other. Historical evidence based on research suggests a 2 percent increase over time, while BLS statistics suggest increases closer to 3 percent. Our

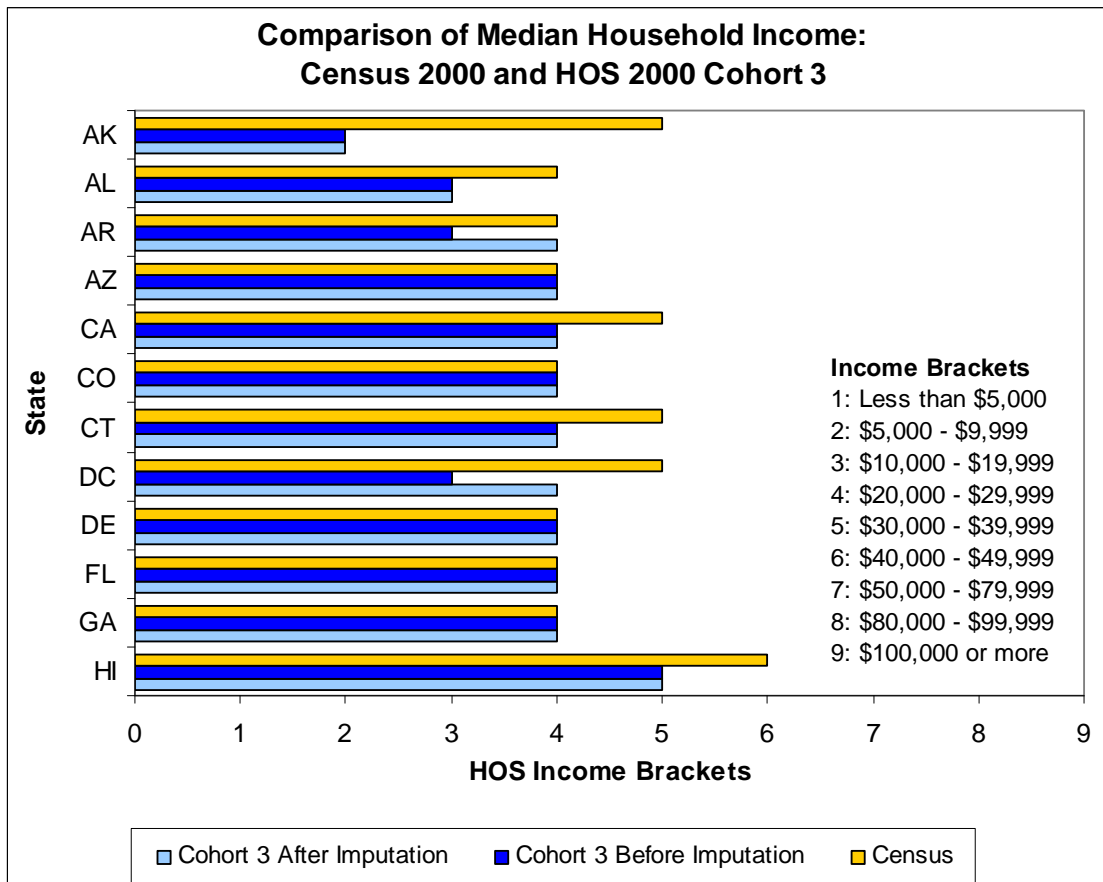
findings do not support an “either—or” decision. The differences in coefficients are too small for us to use one assumption over the other and do not apply to the income variable.

## **5.0 CONCLUSION**

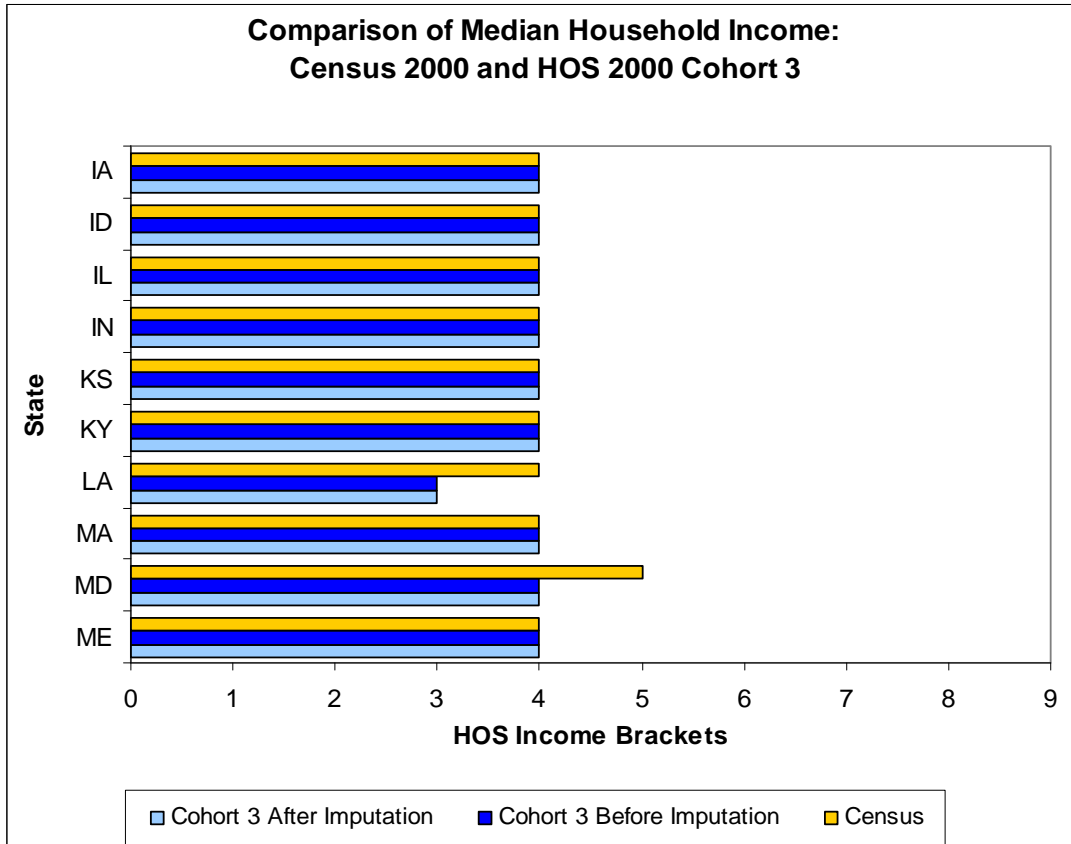
Despite the difficulties in working out the differences between the two data sets and imperfect information for zip code and state-level income that led to different results after imputation, test results for the imputed income variable suggest exploring the methodology further. The coefficients for the income variable did not differ when using a different set of explanatory variables in the three different models. This important finding suggests that imputation methodology does not systematically lead to biased results overall in basic multivariate regression models though the results for respondents with imputed income differ from the larger sample that has income information. Based on our findings, a strong recommendation for a 2 percent or 3 percent inflation adjustment cannot be given, though the underlying imputation methodology seems sufficiently reliable for further analysis or for income imputation in other HOS cohorts. Internal income imputation will show whether these results can be validated further.

## 6.0 FIGURES

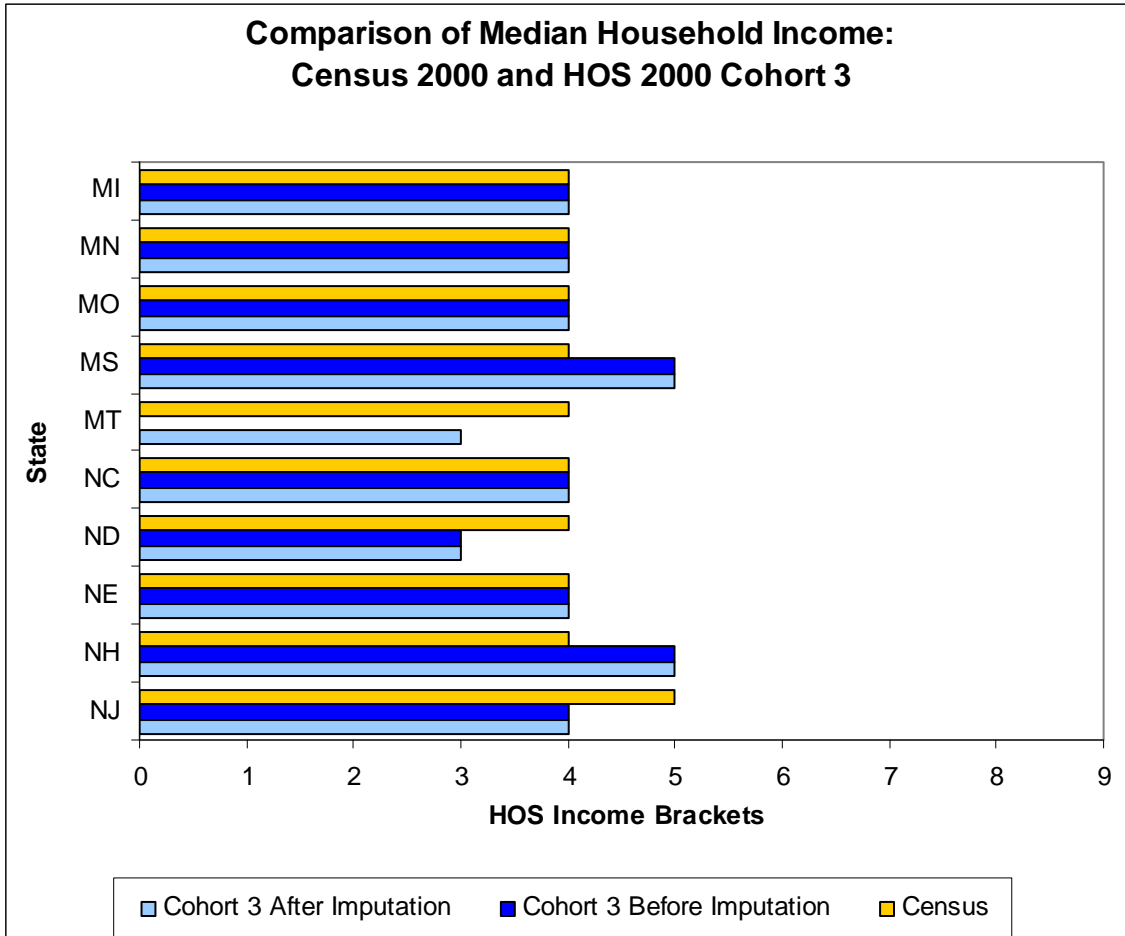
**Diagram 1: Comparison of Median Household Income Census 2000 and HOS 2000 Cohort 3 (Alaska-Hawaii)**



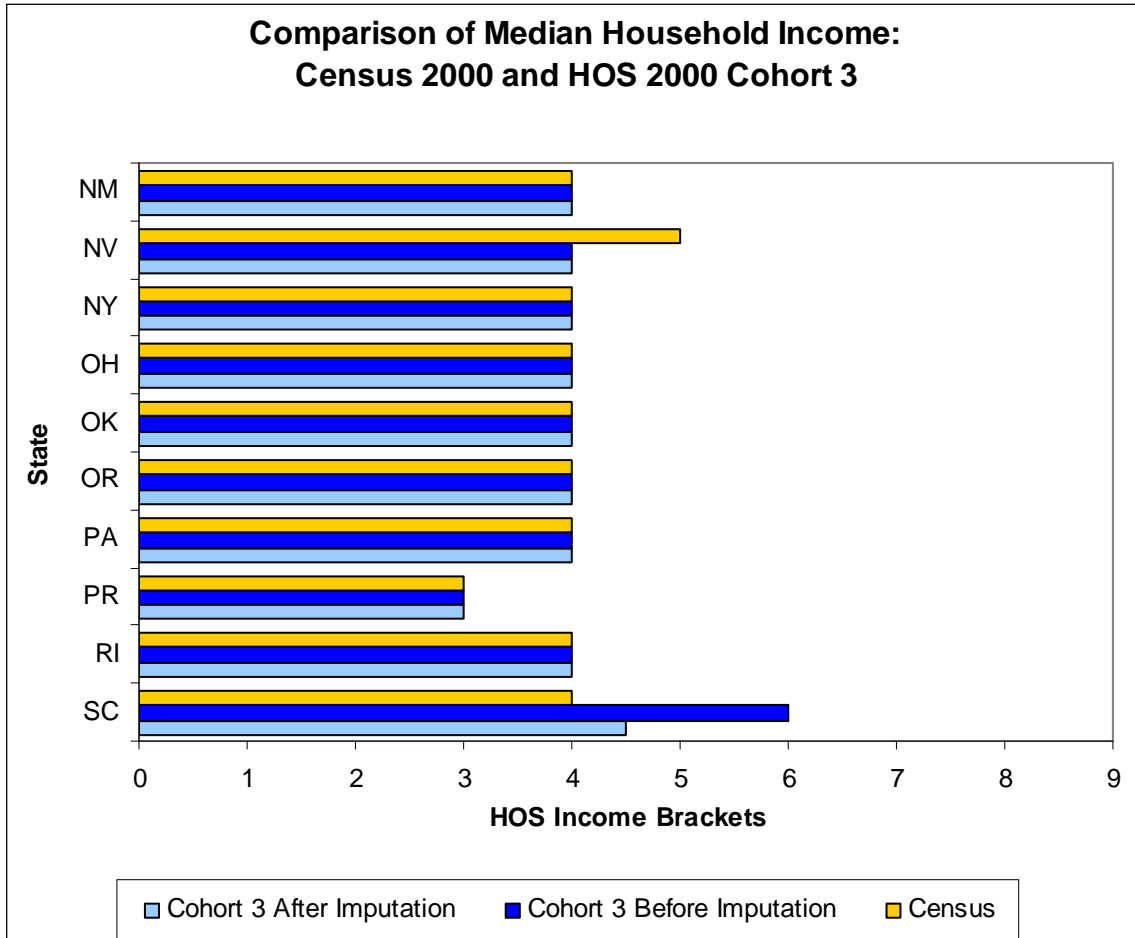
**Diagram 2: Comparison of Median Household Income: Census 2000 and HOS 2000 Cohort 3 (Iowa-Maine)**



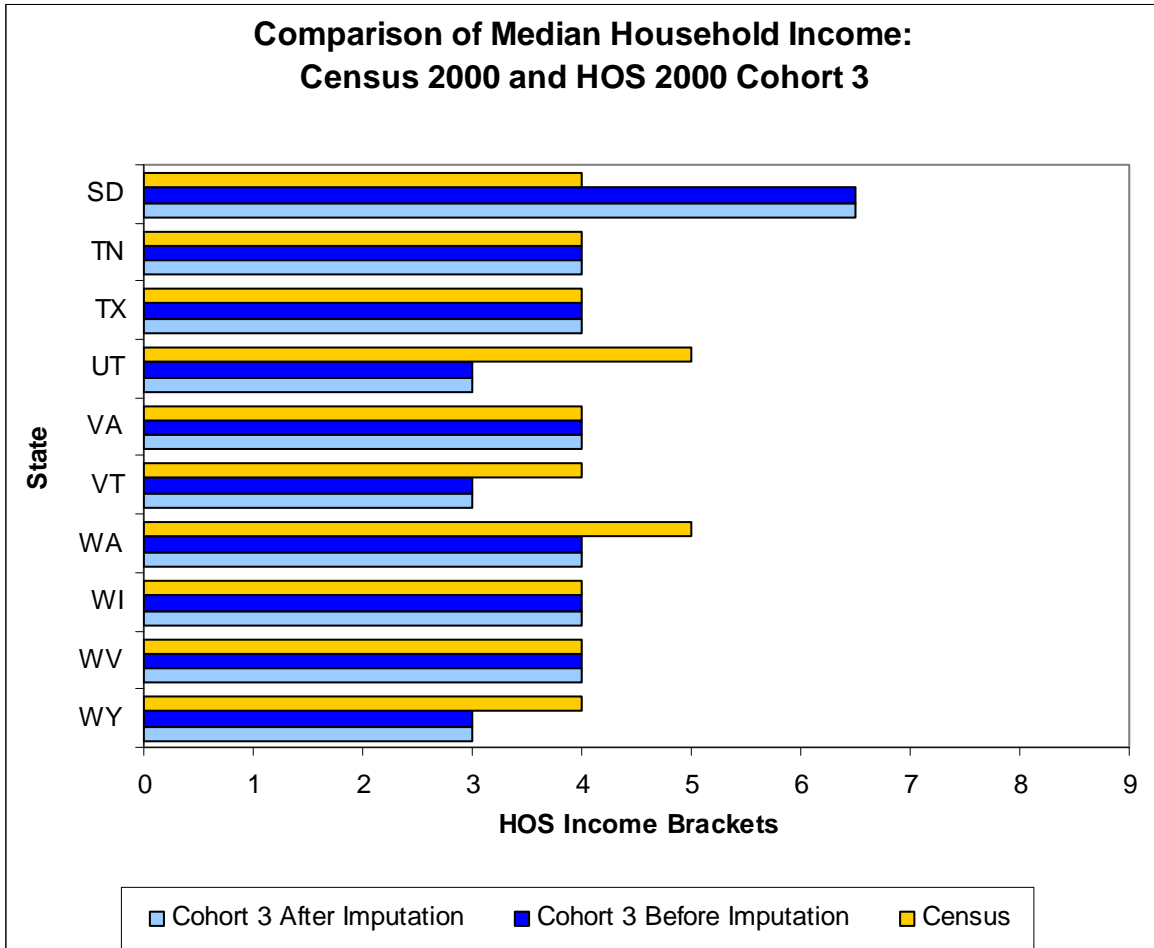
**Diagram 3: Comparison of Median Household Income: Census 2000 and HOS 2000 Cohort 3 (Michigan-New Jersey)**



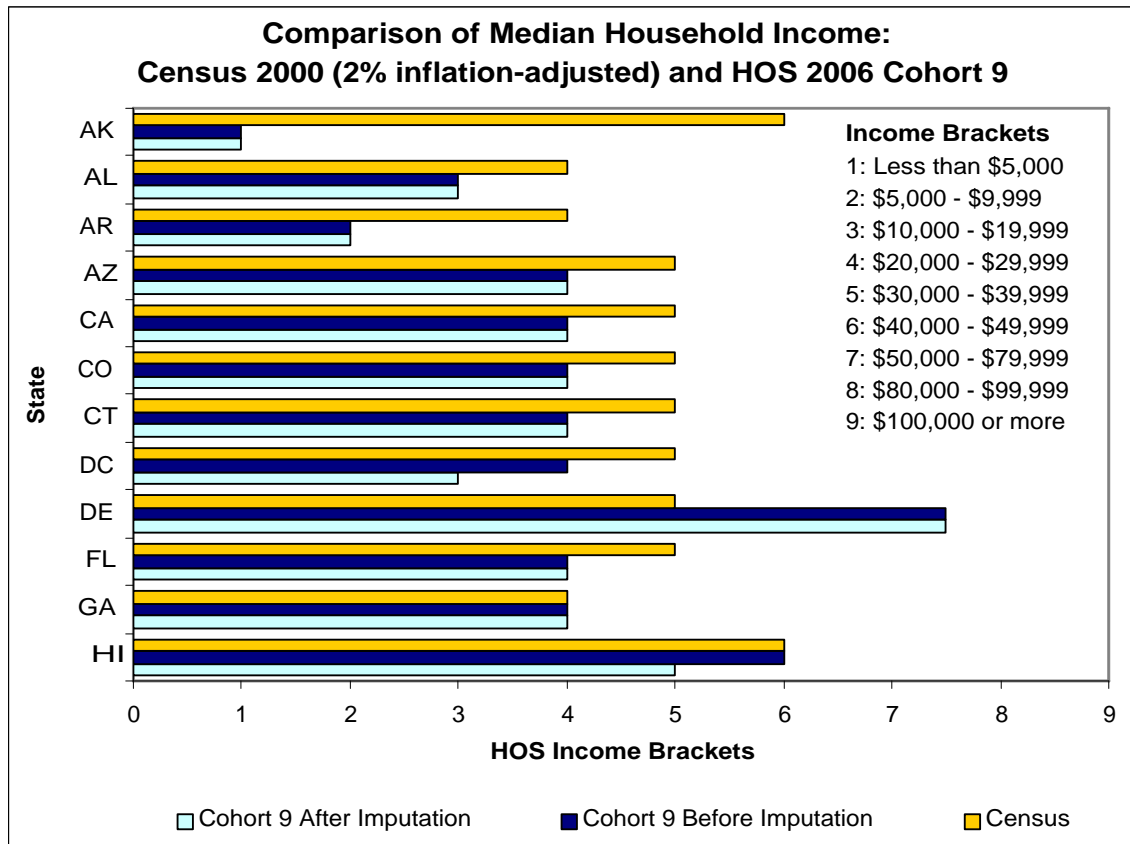
**Diagram 4: Comparison of Median Household Income: Census 2000 and HOS 2000 Cohort 3 (New Mexico-South Carolina)**



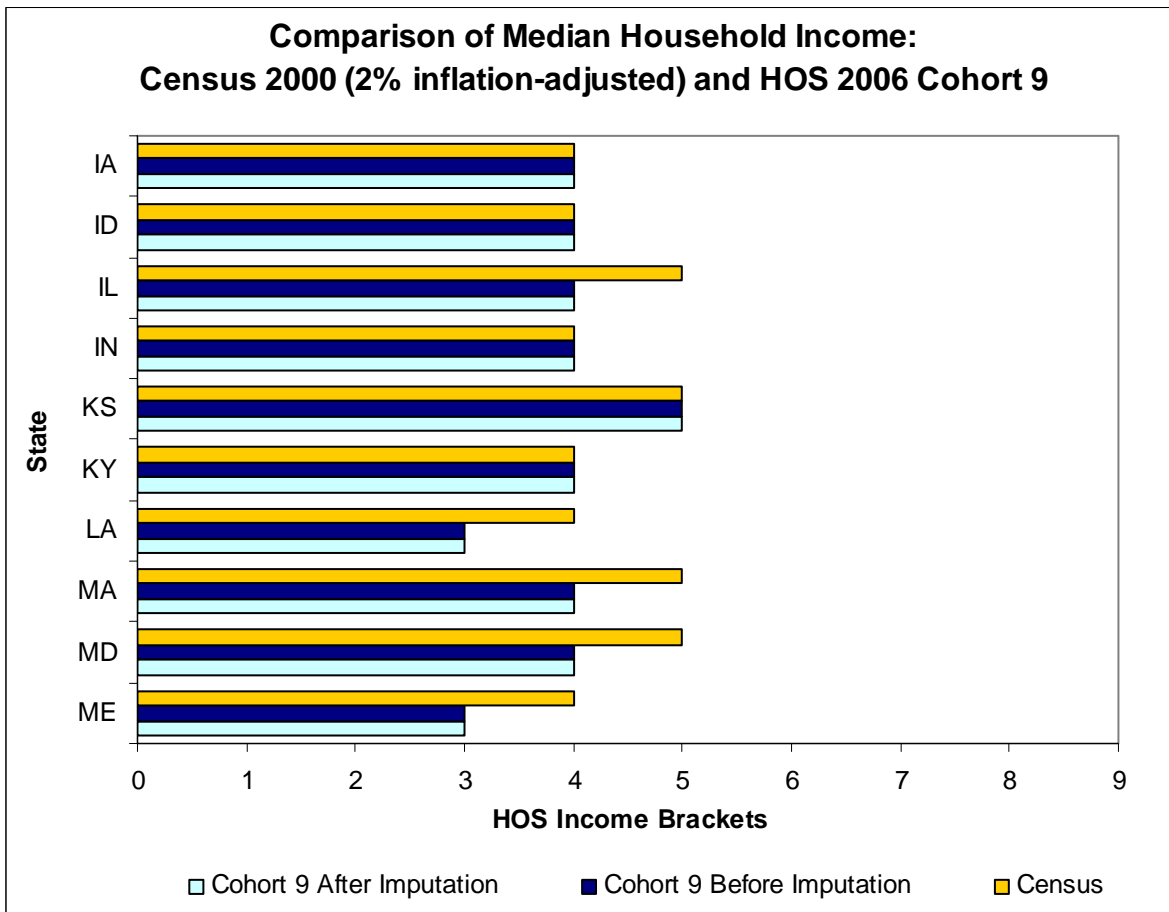
**Diagram 5: Comparison of Median Household Income: Census 2000 and HOS 2000 Cohort 3 (South Dakota-Wyoming)**



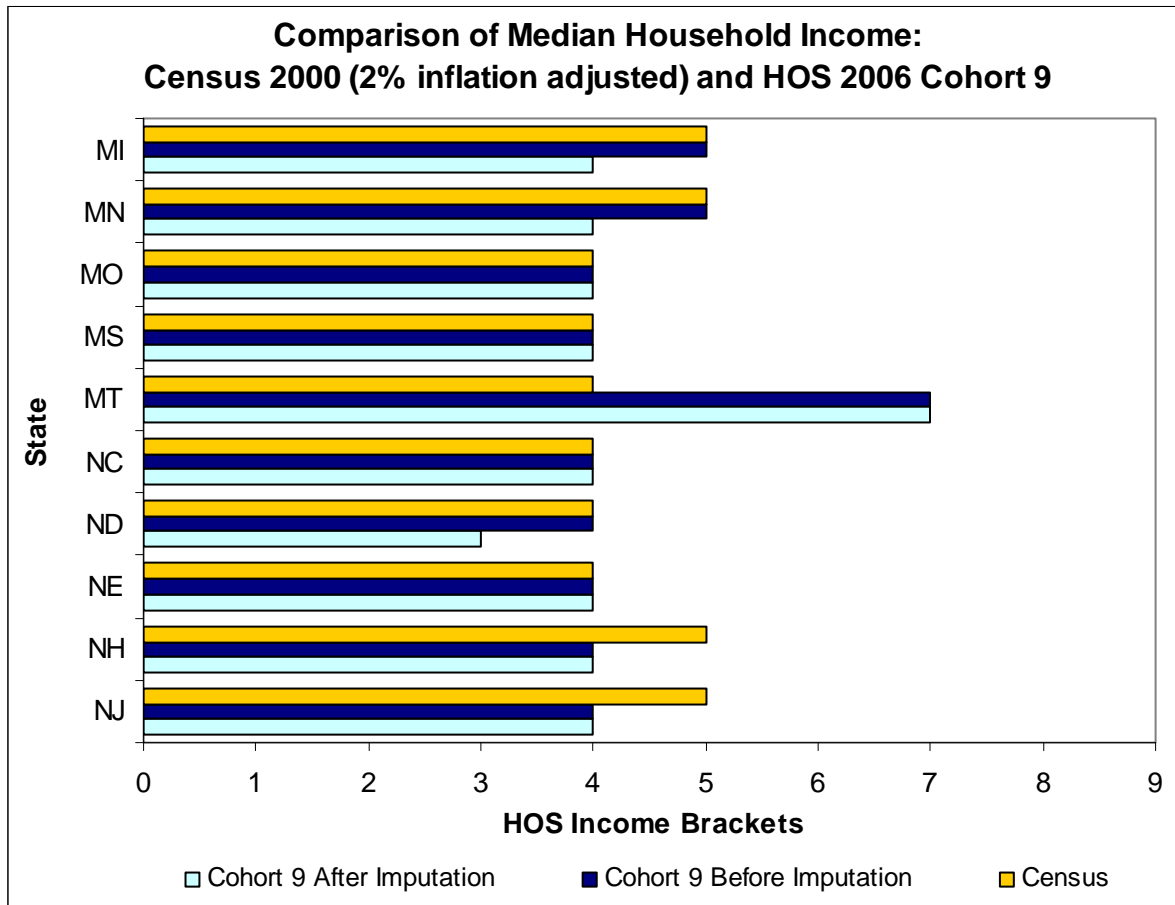
**Diagram 6: Comparison of Median Household Income: Census 2000 (2% inflation-adjusted) and HOS 2006 Cohort 9 (Alaska-Hawaii)**



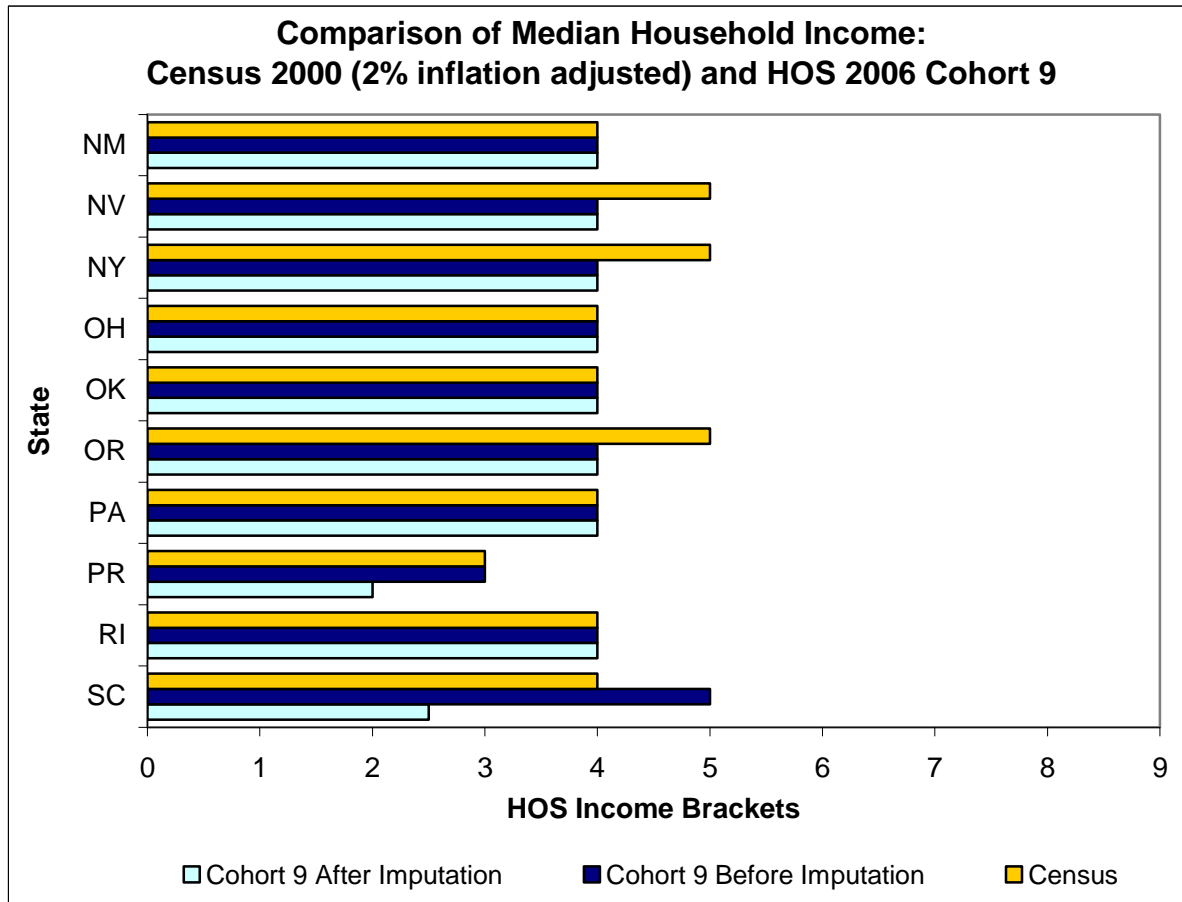
**Diagram 7: Comparison of Median Household Income: Census 2000 (2% inflation-adjusted) and HOS 2006 Cohort 9 (Iowa-Maine)**



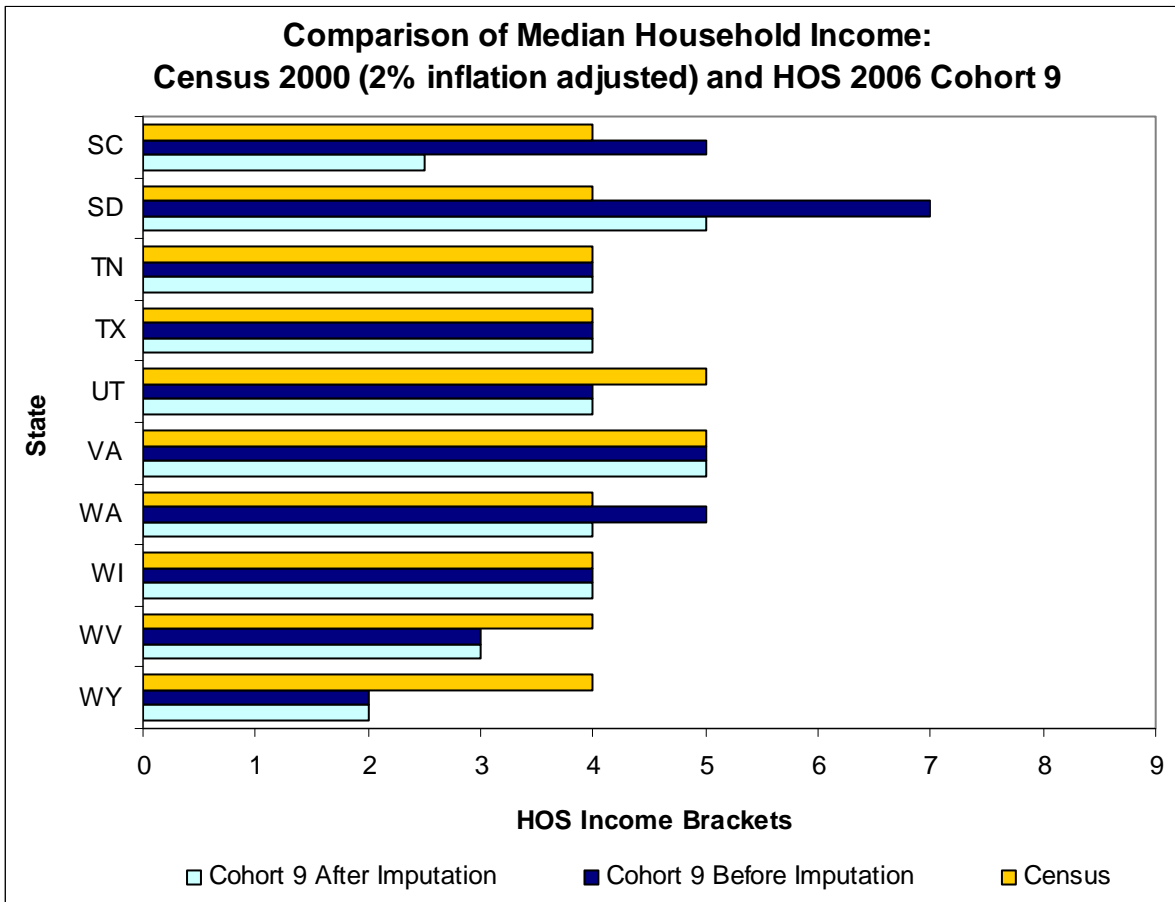
**Diagram 8: Comparison of Median Household Income: Census 2000 (2% inflation-adjusted) and HOS 2006 Cohort 9 (Michigan-New Jersey)**



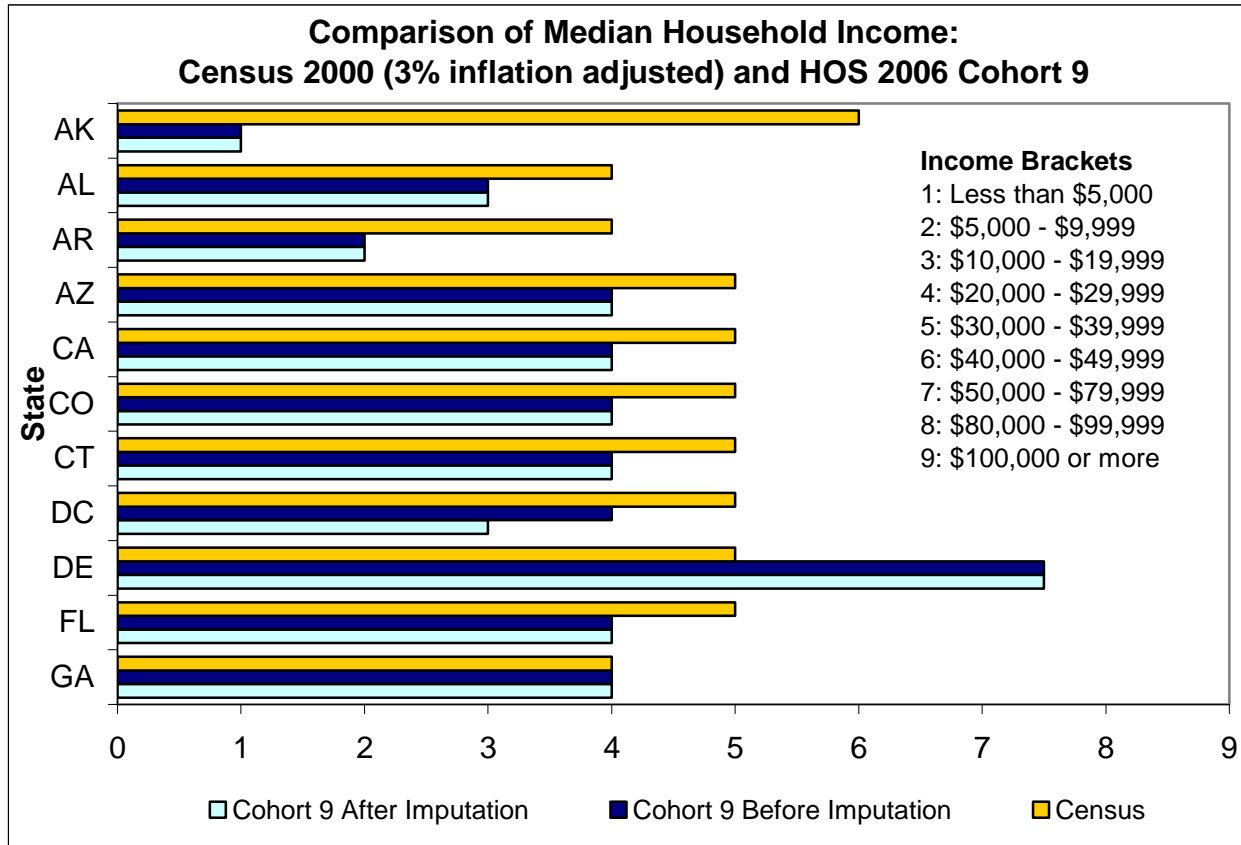
**Diagram 9: Comparison of Median Household Income: Census 2000 (2% inflation-adjusted) and HOS 2006 Cohort 9 (New Mexico-South Carolina)**



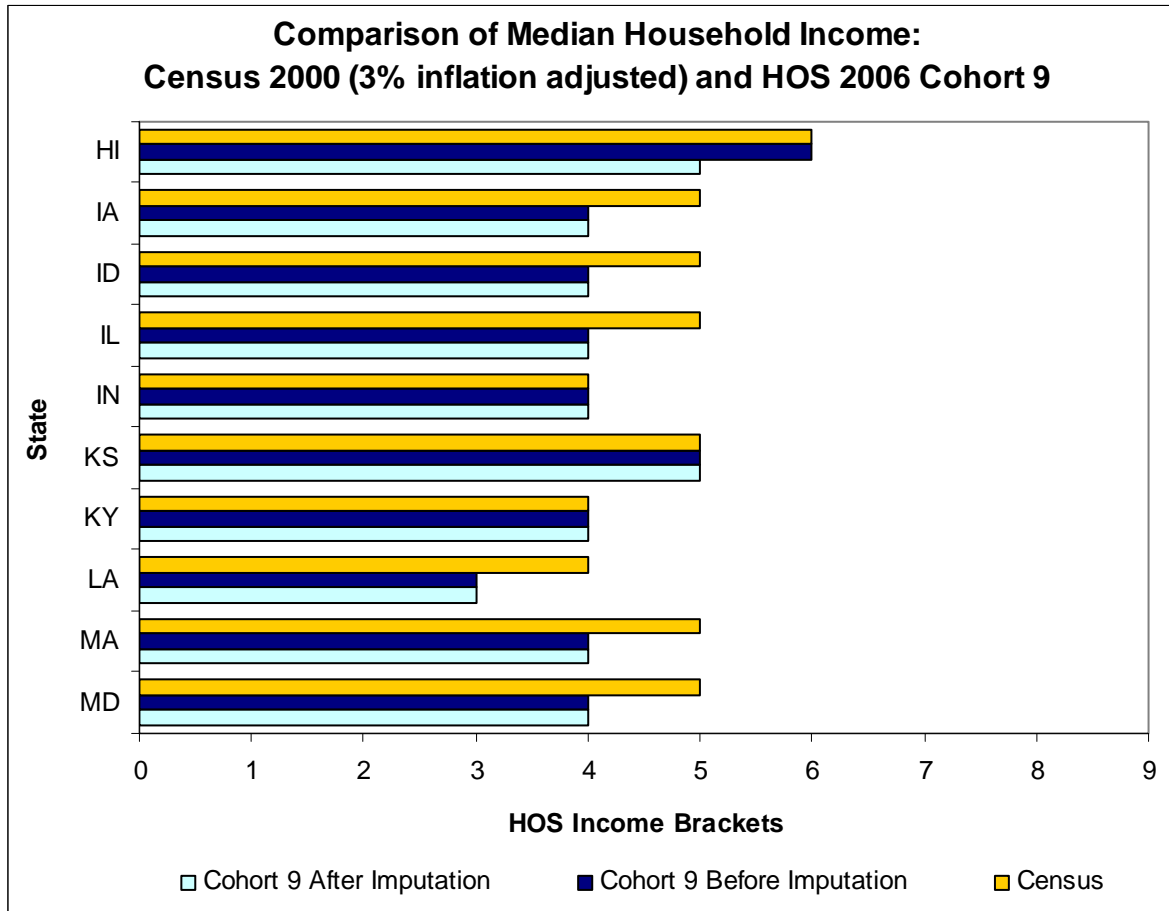
**Diagram 10: Comparison of Median Household Income: Census 2000 (2% inflation- adjusted) and HOS Cohort 9 (South Carolina-Wyoming)**



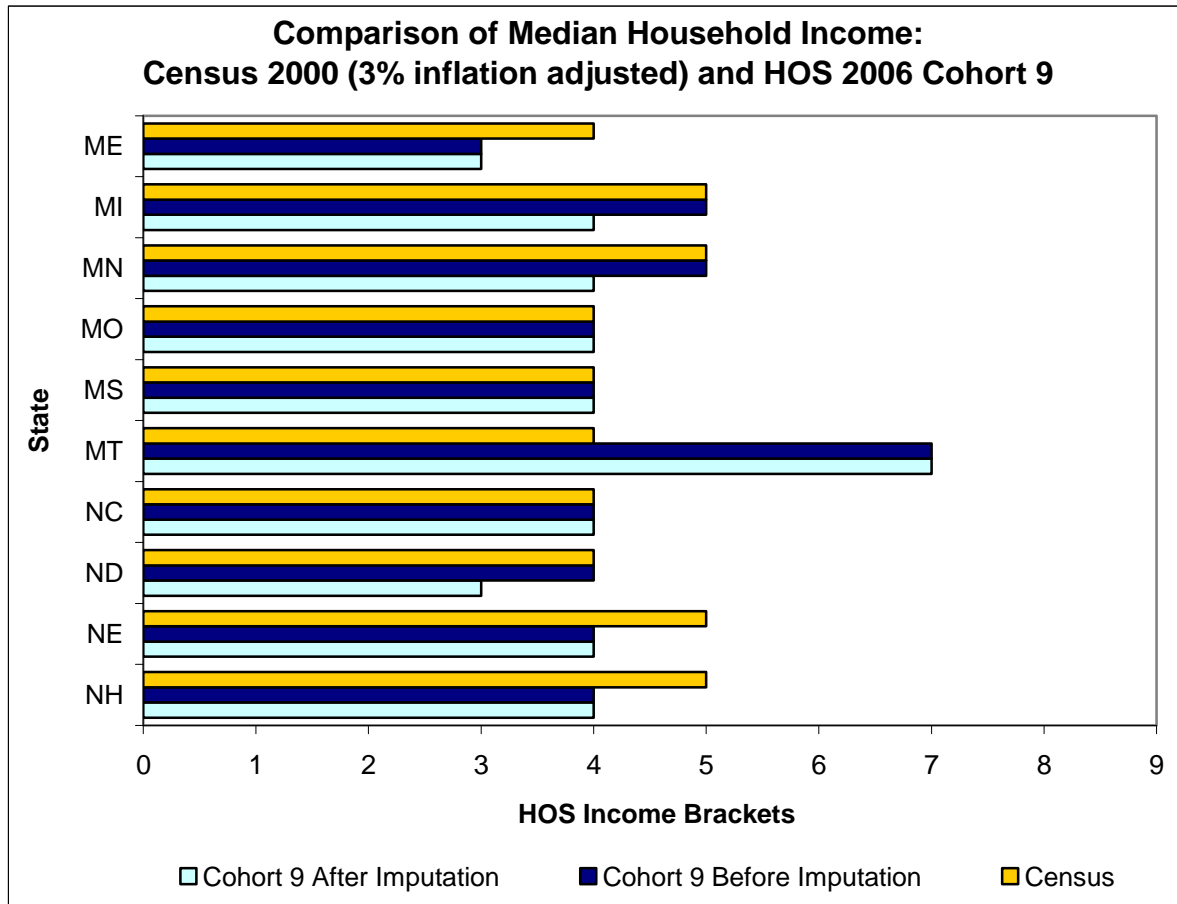
**Diagram 11: Comparison of Median Household Income: Census 2000 (3% inflation-adjusted) and HOS 2006 Cohort 9 (Alaska-Georgia)**



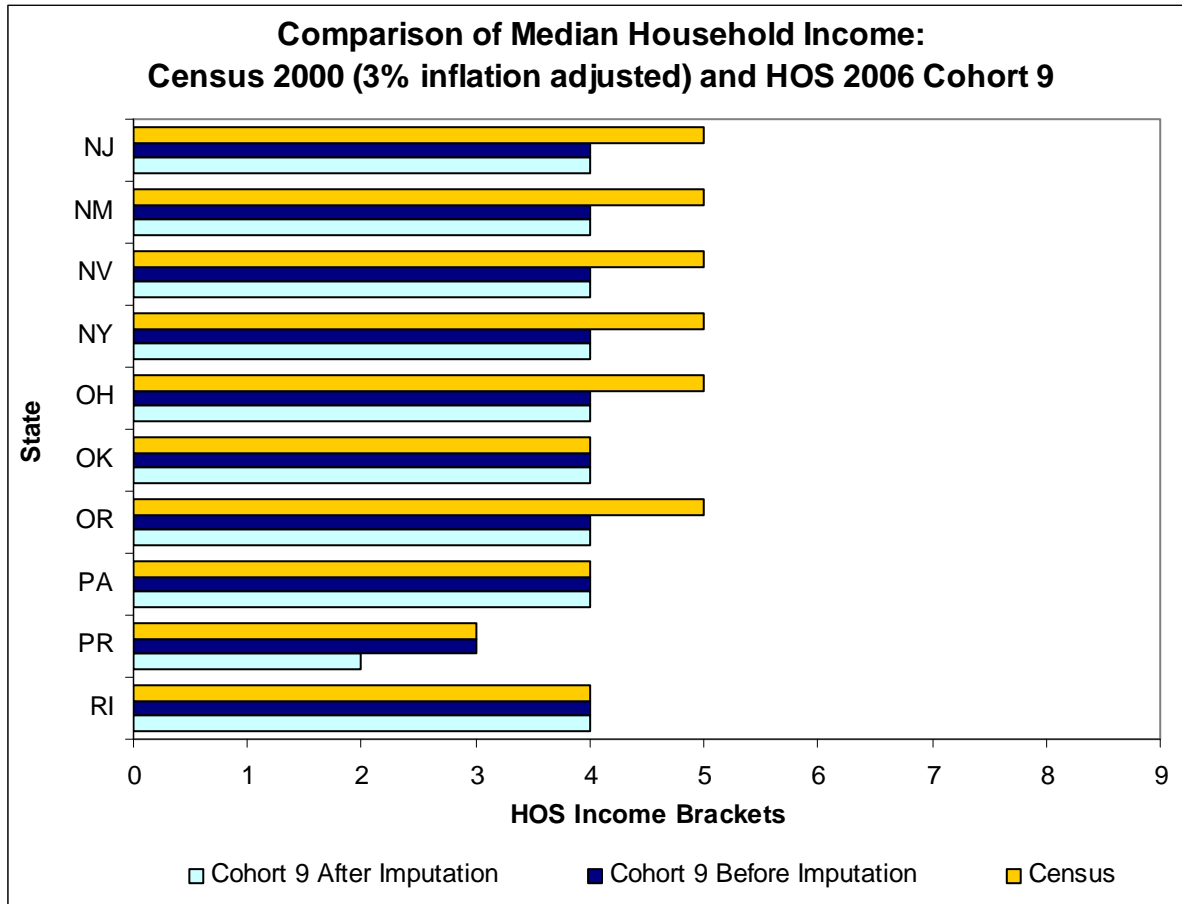
**Diagram 12: Comparison of Median Household Income: Census 2000 (3% inflation-adjusted) and HOS 2006 Cohort 9) (Hawaii-Maryland)**



**Diagram 13: Comparison of Median Household Income: Census 2000 (3% inflation-adjusted) and HOS 2006 Cohort 9 (Maine-New Hampshire)**



**Diagram 14: Comparison of Median Household Income: Census 2000 (3% inflation-adjusted) and HOS 2006 Cohort (New Jersey-Rhode Island)**



**Diagram 15: Comparison of Median Household Income: Census 2000 (3% inflation-adjusted) and HOS Cohort 9 (South Carolina-Wyoming)**

