AN EVALUATION OF A MODEL USED TO OVERSAMPLE LOW INCOME HOUSEHOLDS IN THE 1997 MEPS[1]

Lap-Ming Wun, Steven B. Cohen, John Moeller, Agency for Healthcare Research and Quality (AHRQ)
Lap-Ming Wun, AHRQ, 2101 E. Jefferson Street, Room 500, Rockville, MD 20852

KEY WORDS: logistic regression, predicted probability, low income, poverty.

INTRODUCTION

The Medical Expenditure Panel Survey (MEPS) is conducted by the Agency for Healthcare Research and Quality (AHRQ) and co-sponsored by the National Center for Health Statistics (NCHS). It is conducted to provide nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. It comprises three component surveys: the Household Component (HC), the Medical Provider Component (MPC), and the Insurance Component (IC).

The MEPS HC is the core survey, and it forms the basis for the MPC sample and part of the IC sample. The HC is a nationally representative survey of the U.S. civilian noninstitutionalized population which collects medical expenditure data at both the person and household levels. The HC uses an overlapping panel design in which data are collected through a preliminary contact followed by a series of four rounds of interviews over a 2 and ½ - year period. Using computer-assisted personal interviewing (CAPI) technology, data on medical expenditures and use for 2 calendar years are collected from each household. This series of data collection rounds is launched each subsequent year on a new sample of households to provide overlapping panels of survey data and, when combined with other ongoing panels, will provide continuous and current estimates of health care expenditures. The sampling frame for the MEPS HC is drawn from respondents to the National Health Interview Survey (NHIS) conducted by NCHS. NHIS provides a nationally representative sample of the U.S. civilian noninstitutionalized population and oversamples Hispanics and blacks.

MEPS is a third in a series of national probability surveys conducted by AHRQ on the financing and use of medical care in the United States. The National Medical Care Expenditure Survey (NMCES) was conducted in 1977, while the National Medical Expenditure Survey (NMES) was conducted in 1987. Beginning in 1996, MEPS continues this series with design enhancements and efficiencies that provide a more current data resource to capture the changing dynamics of the health care delivery and insurance system.

The predicting model presented in this report used data from the 1987 NMES and 1996 NHIS to predict the status of sampling units to be selected for the 1997 MEPS.

TARGETED SUBPOPULATIONS TO BE OVERSAMPLED

In addition to the oversampling of Hispanics and blacks inherited from the NHIS, the sample of the HC of the 1997 MEPS was designed to oversample some additional subpopulations of analytic interest. The unit of interviewing and subsampling was the household. To facilitate the sample selection of the new 1997 MEPS sample, the 1996 NHIS households were selected on the basis of the characteristics of the persons they included. There were seven sample domains of interest. An NHIS dwelling unit was assigned to one or more sample domains based on having at least one household member with the characteristic of interest. These sampling domains are not mutually exclusive, their order reflects the hierarchy of their sampling priority. For purposes of sampling, dwelling units containing members having these characteristics were hierarchically classified to form seven mutually exclusive and exhaustive sampling strata. The seven characteristics, in order of sampling priority, are:

1. Adults (age 18 and over) with functional impairments -- at least one activity of daily living (ADL) for which assistance is needed.

[1]The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred.

2. Children (under age 18) with limitations in activity.

3. Individuals 18-64 years old with predicted high medical expenditures.

4. Individuals with family incomes likely to be below 200 percent of poverty level.

5. Adults with other impairments -- ages 18-64, at least one instrumental activity of daily living (IADL) and unable to work; age 70 and over, at least one IADL.

6. Elderly individuals (age 65 and over).

7. All remaining individuals.

Since the 1997 MEPS sample is drawn from the respondents to the 1996 NHIS, a family's 1997 income was not known at the time the sample units were to be selected. Therefore, the status of family income needed to be predicted. In this report, we present the model used to predict the status of family income. (Cohen SB, 2000)

THE PREDICTIVE MODEL

A sample unit's income in a given year would be a reasonable predictor of its poverty status in the next year. However, studies reported in Moeller and Mathiowetz's article (1994) have shown the previous year's reported income on a screener interview is not a very reliable predictor for survey year's poverty status due to under reporting and the dynamic of individuals moving into and out of the poverty in adjacent years. Therefore, Moeller and Mathiowetz (1994) developed a predictive model based on the economic concept of permanent income, which is the family's expected income in a given year based on its human capital and other characteristics and resources. The model was estimated with data from the 1987 NMES and a screener interview conducted in 1986. A slightly modified version of the permanent income model identified the following variables as significant predictors of income status for MEPS sampling purpose:

1. Age of reference person - it is classified into 4 categories: 15 to 25, 25 to 40, 40 to 55, and all others (i.e., less than 15 or older than 55) for the MEPS predictive model.

2. Home ownership.

3. Reporting unit (RU) size.

4. Whether children of specific ages (under 6, 6-15) are present in the RU.

5. Whether someone in the RU other than the reference person is at least 65 years of age.

6. Health status of reference person.

7. Race/ethnicity of reference person - an indicator to identify whether the reference person is Hispanic, black-nonHispanic, or other.

8. Census Division.

9. Metropolitan statistical area (MSA) status and size of the primary sample unit (PSU).

10. Education of reference person - with categories of: no high school, some high school, high school graduate, some college, college graduate, graduate school.

11. Marital status and sex of reference person.

12. Whether reference person or spouse was employed in the previous 3 months.

13. Whether the screener family income of the reporting unit in the prior 12 months was less than 1.25 times the poverty level - In the calculation of the predictive probability for sampling, the value used for this variable is family income in the 1996 NHIS. Poverty status is based on the combination of number of members in the family, family income, and age of family head for one and two-person units.

14. Whether anyone in the RU was covered by Medicaid.

Using these variables as predictors and the poverty status classification as the dependent variable, a logistic regression model was developed for sampling purpose in MEPS to estimate the probability that a reporting unit would have a family income less than 1.25 times the poverty level in a subsequent year. Households with predicted probabilities above a certain threshold value were to be over sampled in the 1997 MEPS. In addition to

facilitating an oversample of individuals with family income less than 125 percent of the poverty level, use of this prediction model was expected to facilitate an oversample of individuals with family income less than 200 percent of the poverty level. Using the data from the 1987 NMES to examine the efficiency of various cut points as the threshold, it was determined that 0.3 was an optimal point in terms of the trade off between maximizing the sample yield and the accuracy of targeting the low income population. Consequently, all reporting units with a predicted probability of .3 or greater to have family income below 125 percent of poverty were oversampled with the expectation of producing a high yield of sampled families with family income less than 200 percent of the poverty level.

The unit of analysis for the permanent income logistic regression model was the reporting unit (RU) (Cohen SB, 2000; Moeller and Mathiowetz, 1994). Estimates of the coefficients of the model were obtained using data from the 1987 NMES and the 1986 screener interview. After the equation was estimated, the 1996 NHIS data were used to calculate logit values. The logit value was then converted to a probability value for each NHIS RU, from which the 1997 MEPS sample was to be drawn, through the equation:

PROB = EXP(LOGIT) / (1 + EXP(LOGIT));

This was the predicted probability that the sample unit would have family income less than 125 percent of the poverty level in 1997. (Moore, 1997)

EVALUATION OF THE PERFORMANCE OF THE MODEL

As described in the last section, the model was built using data from the 1987 NMES (along with a 1986 screener interview) and used 1996 NHIS data to predict the income status of reporting units in 1997. A reporting unit was considered likely to have low income in 1997 if its predicted probability using 1996 NHIS data was greater than or equal to 30 percent. Here, low income was defined as income less than 200 percent of the poverty level. The evaluation is done at the person level because each person in a family has the same poverty status, and the targeted sample yields for the 1997 MEPS were person-based. Based on these criteria, the following weighted sample results are observed:

- Among the 15.4 percent of individuals in families predicted to have low income (i.e., the predicted probability is greater than or equal to 30 percent) the prediction rate of true positives for low income (i.e., income in 1997 was indeed less than 200 percent of the poverty level) is 86 percent.

- Among the 84.6 percent of individuals in families predicted to have high income, the prediction rate for true positives for high income (i.e., income in 1997 was greater than 200 percent of the poverty level) is 75.3.

- Among the 34.2 percent of individuals with income in 1997 less than 200 percent of poverty level, 38.8 percent had been predicted to have low income. Alternatively, among the 65.8 percent of individuals with income in 1997 at or above 200 percent of the poverty level, 96.7 percent had been predicted to have high income.

Further assessment of the performance of the model is given in tables 1 to 3. These tables show the percentage of persons in families with actual low income in 1997 within each of 10 levels of predicted probabilities that the person would be in a low income family in 1997. For example, in table 1, the last column (the last column from left) of both the bar chart and the data table shows that for those individuals in families with predicted probabilities greater than or equal to 90 percent of having income below 125 percent of the poverty threshold, 90.06 percent of them indeed have income less than 100 percent of the poverty level. On the other hand, the first column in table 1 shows that among those persons in families with predicted probabilities not greater than 10 percent, only 5.55 percent are indeed low income. Table 1 shows the distributions with low income set at less than 100 percent of poverty level, table 2 are the distributions with low income set at less than 125 percent of the poverty level, and table 3 has the low income level set less than 200 percent of the poverty level. The desirable result is a high predicted probability of being low income coinciding with the actual income level being low. The first and last columns of table 1 used in the aforementioned example showed that the results are indeed desirable. That is, a large proportion of persons in families who were predicted to have low income (i.e., with high predicted probability) are actually in families with low income, and a large proportion of persons in families with low predicted

probabilities of being low income are in families that are not low income. All three tables show the model performs in a desirable direction. Comparing the three tables, table 3 shows that using low income set at 200 percent of the poverty level as the sampling target, the cut point of 30 percent and above captured a substantially larger proportion of low income units then using other cut point levels with a relatively low percentage of false positives. These results further validate our selection of the 30 percent level for determining our sample selection of low income units, and have shown that the model predicted the poverty status quite accurately.

SUMMARY AND DISCUSSION

In this report we gave a brief introduction to the 1997 MEPS and its oversampling of subpopulations of analytic interest. The subpopulation with low income was one of the domains to be oversampled. However, poverty status of families in 1997 is not known at the time the sample is selected. Therefore, it needs to be predicted. The income level reported in a given year should be an intuitively reasonable predictor for the poverty status in the following year. Studies have shown that due to problems of under reporting of previous year's income on a screener interview and the dynamics of poverty status in adjacent year, the reported income in a given year is not a reliable predictor of the poverty status of the following year. A slightly modified version of a logistic model developed by Moeller and Mathiowetz (1994) based on the concept of permanent income was used to predict the probability of a sample unit's poverty status in the following year. The model coefficients were estimated using data from the 1987 NMES and the 1986 screener interview. The predictive probabilities were calculated using the 1996 NHIS data with the model. Evaluation of the model results by comparing the predictive probabilities and the actual poverty status of sample units in the 1997 MEPS have shown a very high proportion of correct predictions. This result shows that the model is a reliable predictor for poverty status in the next year.

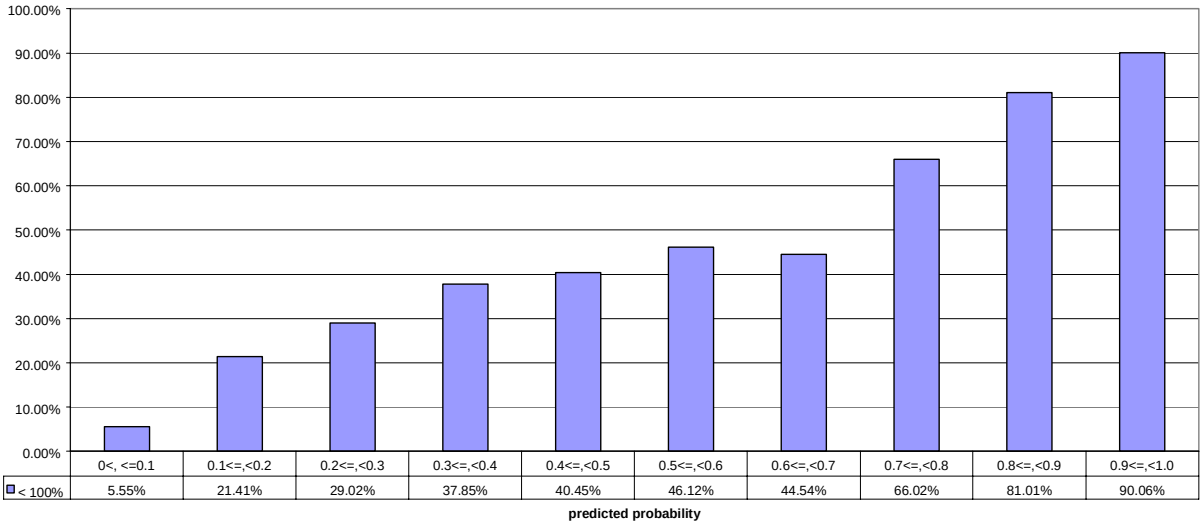
REFERENCES

Cohen SB. Sample design of the 1997 Medical Expenditure Panel Survey Household Component. Rockville (MD): Agency for Healthcare Research and Quality; 2000. MEPS Methodology Report No. 11. AHRQ Pub. No. 01-0001.

Moeller J, Mathiowetz N. "Problems of Screening for Poverty Status", Journal of Official Statistics, 1994, Vol. 10, No. 1, pp. 327-337.

Moore G. Identification of the MEPS 1997 sample from the 1996 NHIS. Bethesda (MD): Social and Scientific Systems, Inc.; 1997. Task NMS2.112 Report.

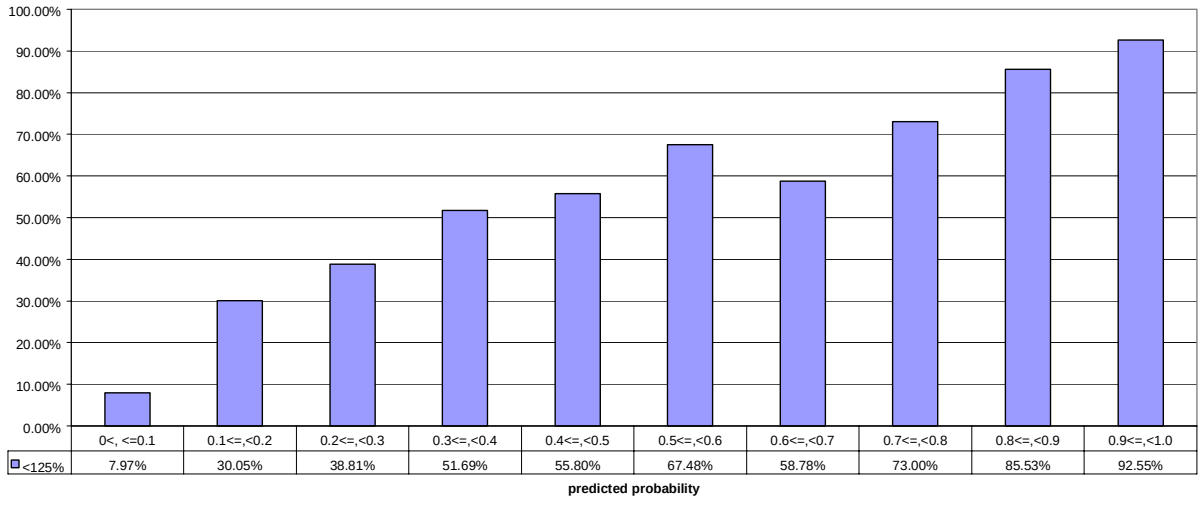
Table 1.
Percent of Sampled Persons in Families with Income Below 100 Percent of the Poverty Level by Predicted Probabilities of Family Income Below 125 Percent of the Poverty level



	0<, <=0.1	0.1<=,<0.2	0.2<=,<0.3	0.3<=,<0.4	0.4<=,<0.5	0.5<=,<0.6	0.6<=,<0.7	0.7<=,<0.8	0.8<=,<0.9	0.9<=,<1.0
< 100%	5.55%	21.41%	29.02%	37.85%	40.45%	46.12%	44.54%	66.02%	81.01%	90.06%

predicted probability

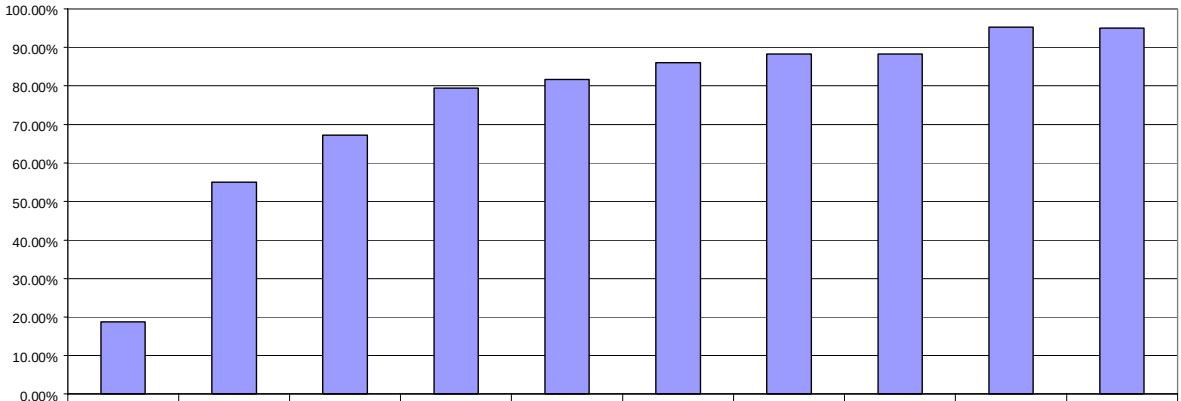
Table 2.
Percent of Sampled Persons in Families with Income Below 125 Percent of the Poverty Level by Predicted Probabilities of Family Income Below 125 Percent of the Poverty level



	0<, <=0.1	0.1<=,<0.2	0.2<=,<0.3	0.3<=,<0.4	0.4<=,<0.5	0.5<=,<0.6	0.6<=,<0.7	0.7<=,<0.8	0.8<=,<0.9	0.9<=,<1.0
<125%	7.97%	30.05%	38.81%	51.69%	55.80%	67.48%	58.78%	73.00%	85.53%	92.55%

predicted probability

Table 3.
Percent of Sampled Persons in Families with Income Below 200 Percent of the Poverty Level by Predicted Probabilities of Family Income Below 125 Percent of the Poverty level



	0<, <=0.1	0.1<=,<0.2	0.2<=,<0.3	0.3<=,<0.4	0.4<=,<0.5	0.5<=,<0.6	0.6<=,<0.7	0.7<=,<0.8	0.8<=,<0.9	0.9<=,<1.0
<200%	18.73%	55.03%	67.26%	79.41%	81.70%	86.00%	88.30%	88.31%	95.21%	95.07%

predicted probability