Subsampling for High Expenditures in the Medical Expenditure Panel Survey - Household Component[1]

Robert M. Baskin[1], Lap-Ming Wun[1]
[1]AHRQ, 540 Gaither Road, Rockville, MD 20850

Abstract

The Medical Expenditure Panel Survey Household Component is an annual two year panel survey of Households sponsored by the Agency for Healthcare Research and Quality and conducted by Westat. The survey collects data on household characteristics, insurance coverage, healthcare use and expenditures. The survey is conducted in overlapping panels with responding units reporting for five rounds of collection covering a two year period. The current work investigates several options for subsampling for further collection in a way that produces overall unbiased estimates but optimizes estimates for high expenditure cases. The possibility of collecting further rounds of data would facilitate trend analysis. Simulated subsampling was done for four methods of sampling: simple random sampling as baseline; probability proportional to size using propensity for high expenditures as size measure; oversampling of high expenditure cases; and stratified sampling with Neyman allocation based on variance of total expenditures. Results of these simulations indicate that if the subsampling is performed at the person level, then either stratified sampling or probability proportional to size allocation are viable options. However, if the subsampling is at the Dwelling Unit level, then stratified sampling with Neyman allocation is clearly optimal.

Key Words: Household Survey, Simulation, Subsampling, panel survey

1. Introduction

The Medical Expenditure Panel Survey – Household Component (MEPS-HC) is an ongoing panel survey of households of the non-institutionalized population of the United States sponsored by the Agency for Healthcare Research and Quality (AHRQ) in coordination with the National Center for Health Statistics (NCHS). MEPS-HC employs a complex survey design which is a stratified multi-stage design with unequal probability selection within strata. Households are subsampled from the eligible responding households in the National Health Interview Survey (NHIS) conducted by the NCHS. Each year of the NHIS becomes a panel of MEPS-HC. The households subsampled from the NHIS are followed for five consecutive rounds covering a two year period so that the MEPS-HC is an overlapping panel design which always has two panels in collection simultaneously. The data are collected through personal household

[1] The views expressed in this paper are solely those of the authors' and do not reflect policy of the Agency for Healthcare Research and Quality nor the Department of Health and Human Services

visits using computer assisted personal interviewing (CAPI). Among the main purposes of the MEPS-HC are to collect data on insurance coverage, healthcare utilization and medical expenses for persons in the U.S. civilian non-institutionalized population.

Both annual and longitudinal data are available from the MEPS-HC. The five rounds of data collection over a two year period are sufficient to measure trends for common events but for relatively rare events, which for many people would include large expenditure items such as inpatient hospital stays, the time frame of two years may not be sufficient. It would be desirable to extend the collection period of the MEPS-HC to three years in order to capture more information on higher expenditure cases. However, extension of the full sample for an additional year of collection would be costly and funding is not currently available for such an extension.

In order to extend collection for one more year while reducing costs an approach of subsampling the second year responders is a possibility but it is desirable to retain as many high expenditure individuals as possible while simultaneously having an unbiased national estimate. The current work looks at various forms of subsampling responders in an unbiased fashion while attempting to maximize accuracy of high expenditure estimates. In order to evaluate this concept multiple simulations are carried out. Since there is no third year of data available for evaluation the approach is to subsample the first year responders using various unbiased methods and evaluate the resulting subsampled estimates in comparison to the full set of second year respondents.

2. Sample Design

As mentioned in the introduction MEPS-HC is an ongoing panel survey of households which are subsampled from the NHIS. Information on the NHIS design can be found in Botman (2000) and information on the MEPS sample design can be found in Ezzati-Rice (2008). The current NHIS design is a multi-stage geographic cluster sample based on information from the 2010 Census[2] and supplemented by a new construction sample. The first level of sampling, referred to as PSUs, is comprised of geographic clusters based on county level information from the Census. The PSUs are stratified within state by population and selected probability proportional to size. Within each PSU the housing units are subdivided into geographic areas called *segments* in a way that each segment has a minimal number housing units. The segments are grouped into strata based Census information plus one stratum for new construction. Within each stratum, segments are selected probability proportional to size with the size measure based on the number of housing units in the segment. For each segment the housing units are listed and selected within segment with equal probability. If a housing

[2] When this study was conducted the latest five years of data were used starting from 2005. The NHIS design began using Census 2010 data in their 2006 sample design so the early years of the MEPS data in this study are based on the 2000 Census data.

unit is deemed eligible then all eligible persons within the household are included in the sample.

The MEPS-HC takes a subsample of the eligible responding units in the previous year of NHIS in MEPS sampling strata based on priority populations. For a list of specific priority populations in given years see Table 1 in Ezzati-Rice (2008). If an NHIS responding unit is selected for the MEPS-HC then AHRQ will attempt to interview the same responding unit, even if the people who have moved from the address at which they were interviewed for the NHIS. This provides NHIS variables for nearly all MEPS-HC respondents. The current MEPS-HC is designed to roster and collect information on all eligible persons within the unit. In the current design the MEPS-HC goes back to the responding units for five rounds of collection. MEPS-HC collects data on medical expenses, health insurance, and healthcare utilization as well as socio-demographic information about the sampled persons.

3. Extending the Sample

The current collection of five rounds of data reflects a compromise between burden and costs and the ability to measure trends. However there is some desire to extend the rounds of collection to obtain more information for trend analysis. One option is to subsample the second year responders and field a subsample for further collection that would produced unbiased national estimates but also sufficiently accurate estimates of the high expenditure cases. Of course a simple random sample would satisfy the unbiased national estimates but because expenditures are positively skewed this would not provide sufficiently accurate estimates of the high expenditure cases. There are alternatives for sampling that would provide unbiased national estimates but could increase the accuracy of the high expenditure cases. There is a method of oversampling that is simple but increases the variance. There is a method of sampling probability proportional to size (pps) sample based on a propensity to have a high expenditure that requires modeling expenditures but in certain cases reduces the variance compared to oversampling. Another method is to stratify the sample based on previous years expenditures and allocate the sample to the strata using an optimal allocation method. This should produce minimal variance estimators.

3.1 Oversample Previous Year High Expenditure Cases

One simple approach is to choose a cutoff for high expenditure and oversample at some level the cases above the cutoff in previous collection. In previous work Moeller (2002) chose the 85th percentile as a cutoff for high expenditure cases. Then all cases in the top 15^{th} percentile are chosen and the rest are sampled using a simple random sample. In the simulations different levels of cutoffs were used but a fixed cutoff of 85^{th} percentile is shown in the tables.

3.2 PPS Sampling

Another approach is to model a propensity for having high expenditures and take a probability proportional to size (pps) sample based on the propensity. The model used to calculate the propensity of high expenditure is based on the model established in Moeller (2002). The dependent variable for the logistic model is an indicator based on the actual expenditures of the first year. The high/low expenditure criterion established in the 2002 study was that an individual's health expenditure is designated as in the "high" category if it fell in the top 15 percent of the distribution of medical expenditures of the population (i.e., the "cutoff" is the 85th percentile). The predictor variables used in this study included all those used in the 2002 model(age, gender, self reported health status, Census region, Metropolitan Statistical Area status, marital status, poverty status, live alone status, presence of health limitations, number of ambulatory visits), and an additional indicator for diabetes status. The probability that the person would incur high medical expenditures as calculated by this logistic model is commonly called the propensity score. If an individual's propensity score falls in the top 15 percent of the distribution of propensity scores of the population, the individual is designated as being in the predicted "high" category. For DU level sampling, as long as there is one person in the DU with predicted high expenditure, this DU is designated as in the "high" category.

3.3 Stratification with Neyman Allocation

A different approach is to stratify based on previous year's expenditure and use Neyman allocation within the strata. In this study the strata boundaries were chosen using the cumulative square root of *f* rule and the allocation is done to minimize the variance of total expenditures. The cumulative square root of f rule and Neyman allocation are explained in Cochran (1977). One decision necessary for stratification is how many strata to use. Section **5A.8** in Cochran (1977) addresses the issue of number of strata. Because of uncertainty of the optimal number of strata, stratifications with 3, 5, 7, and 10 strata were considered for the Neyman approach.

4. A Simulation of Subsampling First Year Cases

Currently no three years of data collection exists for study so all of this investigation is based on two years of data collection. The approach is to use information in the first year collection to subsample the first year units and compare the estimates from the subsample in the second year to the full second year collection. In this study, the full second year represents a "gold standard" to measure effectiveness of the methods of subsampling. This also allows calculation of the variance of the second year subsample conditional on knowing the second year full sample variance (see chapter 12 in Cochran (1977)).

To this end a simulation was run in which each method of subsampling was used to sample either persons or dwelling units(DUs) from the first year sample and then estimates from the second year values of the corresponding subsampled units were compared to the full second year estimate. Since each method of

subsampling was intended to be unbiased, except in the special cases noted below, the point estimates from the subsample should differ from the full second year point estimate by only the simulation error that comes from randomly repeating the sampling. In this case, the sample variances from the subsampled units are used to compare the effectiveness of the methods.

4.1 DU vs. Person Level Subsampling

In the simulation, sampling was done at either the person level or the dwelling unit (DU) level. If the plan is to return to the field to collect the data using the current instrument then sampling DUs makes sense. However, if the third year follow-up is intended to be by telephone then it would be possible to follow individuals and it may not be necessary or desirable to collect data for all individuals in the DU. In one sense the person level subsampling is more natural. The expenditure variables are at the level of the person. In order to use the different sampling methods at the DU level it was necessary to create DU level estimates of expenditures so the expenditures for all individuals in the DU were aggregated to the DU level for DU expenditure variables. In this simulation the variables used were Total Expenditures, Out Patient Expenditures, Office Based Expenditures, and Emergency Room Expenditures.

4.2 Parameters for the Simulation

In the simulation two arbitrary parameters were set. One parameter was the sample rate which was set at levels of 25%, 30%, 35%, 40%, 45%, and 50%. The second parameter, which was needed only for the methods of oversampling and pps, is the cutoff for high expenditure cases. This parameter was set at 85% or 90% or 95% for *individual* total expenditures. Note that if the cutoff for high *individual* total expenditure was set at 85% then approximately 35% of the first year DUs had individuals with total expenditures above this cutoff. Since ultimately it is desired to have an accurate estimate of high *individual* total expenditures then for DU level subsampling the oversample should be of DUs containing any individual having high total expenditures as opposed to the DU level expenditure being above a cutoff. This can create situations in which the oversample method, in order to be unbiased, requires more sample than can be allocated. This can be seen in the means from Table 1. With a cutoff of 85% a sample rate of 25% of the DUs will not produce an unbiased estimate of mean expenditures for the full second year of data (note the bolded mean for total expenditures) but with a sample rate of 40% of the DUs the mean of the total expenditures from the oversample is less than a tenth of a percent off from the full second year mean of total expenditures.

4.3 Data for the Simulation

The data used for the simulation came from five panels of MEPS-HC data. The five years of data were first fielded in 2005 to 2009. The approach to subsampling described here is to use each panel as a separate frame and creates five simulated results. The results of the five years were then averaged to produce information on the subsampling. The exact magnitude of the sample variances for the methods

was different for the two approaches but the rank of the accuracy of the subsampling methods across the two approaches was the same.

4.4 Simulation Methodology

Given the frame in the separate panel approach, the simulation would sample the first year data based on the methods of simple random sample (SRS), oversample, pps sample, and stratified sample with 3, 5, 7, or 10 strata and then proceed to do the same subsampling for each subsequent year of data. For the stacked panel approach the methods of subsampling were almost the same but because of the length of time for the simulation of the DU level subsampling with stacked data, only stratifications with 5 and 10 strata were used in the comparison.

For each of the approaches and for both person and DU subsampling, the simulation replicated the subsampling 2,000 times. In each replication, the mean of interest for the five expenditure variables was estimated for the overall sample, and for high expenditure subdomains. The means for the full sample are known so the differences between the subsampled mean and the known full sample mean could be used to calculate a bootstrapped 'second phase' variance of the subsampled method. The combination of the full data 'first phase' variance with the bootstrapped 'second phase' variance allows calculation of the variance of the estimator under each method (see Chapter 12 in Cochran (1977)).

5. Results of the Simulation of Subsampling First Year Cases

Except for the noted case of the Oversample method requiring more sample than could be allocated, because if you want to sample all of the high expenditure cases and the allocated sample is smaller than the total number of high expenditure cases, all of the methods were unbiased and the average of the full sample means across the 2,000 simulation replications were usually within .1% of the full sample mean as can be seen in Table 2.. There was a definite ranking among the standard errors for the methods. For DU level sampling the Oversample method and SRS methods had standard errors on average approximately equal across all possible combinations of sample rates and cutoff values followed by the PPS method. The Neyman 5 allocation method was always best with Neyman 3 not a s good as Neyman 5,7, or 10 while Neyman 5,7, and 10 were almost identical. The observation that Neyman allocation with five levels of strata is as good as Neyman allocation with more levels of strata agrees with the general rule of thumb that five strata will capture 95% of the information in a categorized numerical variable as explained in section **5A.8** of Cochran (1977).

If the combinations of sample rates and cutoff values which caused the Oversample method to require more sample than was allowed were eliminated, then the Oversample method was on the same order as the PPS method on average. The results for the DU level sampling at a 40% sample rate can be seen in figure 1. The value in the plot is the ratio of the standard error of the

subsampling method divided by the standard error of the full second year sample for the given variable. There are four domains of estimation in the plot. The first domain is estimating the mean expenditures of the full population. The other three subdomains estimate mean expenditures for people with expenditures above the given percentiles 85%, 90%, and 95%, representing the upper tails of the expenditure distribution for the given variables. There are also three levels of cutoffs in the plot representing using a cutoff for high expenditures for Oversampling and for PPS sampling of 85%, 90%, and 95%. Recall that the Neyman allocation, PPS model, and strata for oversampling were based on Total Expenditures. The results for the three variables Total Expenditures, Office Based Visits, and Out Patient Expenditures indicate that Neyman allocation is the best approach. But for the variable Emergency Room Expenditures the SRS method was better than the Neyman allocation based on Total Expenditures. The method of Oversampling fared poorly in almost every regard for the national estimate while the method of PPS had some promising aspects. Also note that the relative size of the ratio of standard errors across the subdomains since the divisor reflects the increased variance in estimating increasingly smaller subdomains. In the case of using a cutoff of 95% for high expenditures, the Oversample method performed as well as the Neyman method but the higher cutoff made the PPS method worse.

Finally, notice that since the sample rate is 40% then the expected ratio of the standard error of the simple random sample to the full year sample should be approximately 1.58 which is the reciprocal of the square root and this was fairly accurately shown in the figure. A vertical line is placed on each panel of the graph at this point as a reference.

For the person level sampling the results, except for the PPS sampling, were similar while the PPS method improved to the point that it was competitive with the Neyman allocation. The difference must be because the predictive model is more accurate in predicting person level than DU level expenditures.

To this end a simulation was run in which each method of subsampling was used to sample either persons or dwelling units(DUs) from the first year sample and then estimates from the second year values of the corresponding subsampled units were compared to the full second year estimate. Since each method of subsampling was intended to be unbiased, except in the special cases noted below, the point estimates from the subsample should differ from the full second year point estimate by only the simulation error. In this case, the sample variances from the subsampled units are used to compare the effectiveness of the methods.

6. Conclusion

It may be desired in the future to subsample the MEPS-HC based on expenditures. This study provides information to guide that subsample. It is clear that the

Oversample method is the least desirable statistical approach while Neyman allocation with five strata is close to optimal for both person and DU level sampling. In the person level sampling the PPS method also provides a strong approach.

Table1:

Mean Expenditures and SE from 100% Oversampling Compared to Full Second Year Data by Overall Sample Proportion

			Total	Out Patient	Office Based	Emergency Room
Full Data		Mean	$3,734.35	$336.96	$883.18	$130.33
		SE	$122.77	$24.73	$35.61	$7.51
Original Sample Cutoff	Overall Sample %		Total	Out Patient	Office Based	Emergency Room
85%	25%	Mean	$4,204.90	$362.30	$977.96	$138.88
		SE	$3,158.20	$518.72	$811.78	$265.66
	30%	Mean	$4,220.79	$366.71	$989.33	$141.56
		SE	$2,902.26	$570.59	$919.07	$270.35
	35%	Mean	$4,263.21	$371.22	$984.06	$143.23
		SE	$2,999.00	$559.50	$710.36	$315.85
	40%	Mean	$3,737.95	$337.32	$884.16	$130.40
		SE	$243.26	$50.50	$72.28	$21.44
	45%	Mean	$3,736.15	$337.28	$883.11	$130.34
		SE	$180.48	$37.96	$53.43	$15.11
	50%	Mean	$3,733.89	$336.94	$882.91	$130.22
		SE	$159.17	$32.98	$47.30	$12.61
90%	25%	Mean	$4,310.14	$382.89	$995.82	$144.29
		SE	$3,668.87	$718.98	$893.57	$321.03
	30%	Mean	$3,734.87	$336.75	$884.01	$130.25
		SE	$283.66	$55.59	$80.19	$24.21
	35%	Mean	$3,738.87	$337.04	$884.74	$130.58
		SE	$211.74	$41.52	$60.58	$17.25
	40%	Mean	$3,735.99	$336.99	$883.50	$130.32
		SE	$182.62	$36.09	$52.29	$14.18
	45%	Mean	$3,734.83	$336.86	$883.22	$130.59
		SE	$166.65	$32.73	$47.67	$12.56
	50%	Mean	$3,735.54	$337.29	$883.15	$130.54
		SE	$155.49	$30.83	$44.72	$11.44
95%	25%	Mean	$3,737.80	$337.15	$882.90	$130.18

		SE	$232.81	$46.81	$68.65	$18.49
	30%	Mean	$3,736.37	$336.69	$883.45	$130.44
		SE	$202.96	$40.63	$59.70	$15.64
	35%	Mean	$3,736.22	$336.91	$883.36	$130.50
		SE	$183.34	$36.87	$53.88	$13.83
	40%	Mean	$3,734.78	$336.96	$883.38	$130.36
		SE	$170.46	$34.28	$49.93	$12.49
	45%	Mean	$3,735.26	$337.09	$883.17	$130.26
		SE	$160.81	$32.28	$47.17	$11.58
	50%	Mean	$3,733.80	$336.93	$882.95	$130.26
		SE	$152.31	$30.88	$44.84	$10.81

Sampling is at DU level by panel and results from five panels are averaged together.

Table 2.

DU Subsampling of First Year Cases With a 25% Sample Rate and 85% Cutoff for Oversampling				
	Total	Out Patient	Office Based	Emergency Room
truth Mean	$3,734.35	$336.96	$883.18	$130.33
truth SE	$122.77	$24.73	$35.61	$7.51
SRS Mean	$3,735.49	$336.80	$883.72	$130.33
SRS SE	$238.74	$48.48	$69.47	$14.85
Oversample Mean	$4,204.90	$362.30	$977.96	$138.88
Oversample SE	$3,158.20	$518.72	$811.78	$265.66
PPS Mean	$3,738.35	$337.39	$883.99	$130.16
PPS SE	$225.24	$45.60	$64.63	$18.34
Neyman.3 Mean	$3,734.75	$336.93	$883.55	$130.39
Neyman.3 SE	$223.90	$47.65	$65.18	$17.78
Neyman.5 Mean	$3,732.94	$336.62	$883.99	$130.24
Neyman.5 SE	$202.42	$43.15	$59.46	$16.47

Neyman.7 Mean	$3,733.74	$336.79	$883.55	$130.50
Neyman.7 SE	$202.20	$42.67	$58.74	$15.92
Neyman.10 Mean	$3,734.22	$336.51	$882.83	$130.27
Neyman.10 SE	$202.22	$42.78	$58.16	$16.07

Note: 'Truth' is the full second year sample information Sampling is by panel and results of five panels are averaged together.

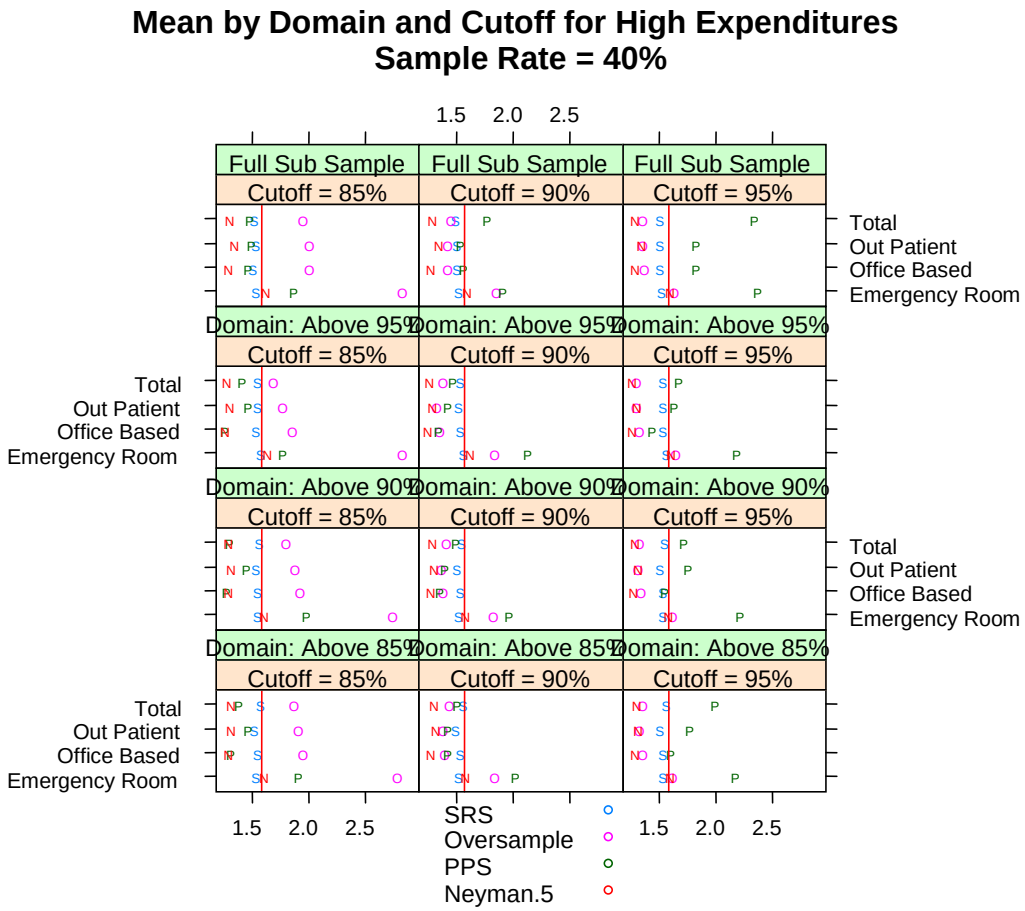

Figure 1: Ratio of Standard Errors of Subsampling Methods to Standard Error of Full Second Year Sample, sample rate = 40%.

Acknowledgements

The authors would like to thank Steve Cohen, Steve Machlin Sadeq Chowdhury, and Joel Cohen for support and helpful comments

References

S. L. Botman, T. F. Moore, C. L. Moriarity, and V. L. Parsons, *Design and estimation for the National Health Interview Survey, 1995–2004*, National Center for Health Statistics. Vital Health Stat **2(130)** (2000), 645--646

T. Ezzati-Rice, F. Rohde, J Greenblatt, *Sample Design of the Medical Expenditure Panel Survey Household Component, 1998 - 2007*, Agency for Healthcare Research and Quality, Rockville, MD, Methodology Report No. **22**, (2008)

John F. Moeller, Steven B. Cohen, Nancy A. Mathiowetz, and Lap-Ming Wun, *Model-based sampling for persons with high health expenditures: evaluating accuracy and yield with the 1997 MEPS*, ASA Proceedings of the Joint Statistical Meetings, (2002), 2367-2372, American Statistical Association (Alexandria, VA)

William G. Cochran, *Sampling Techniques*, J. Wiley and Sons, New York, 3rd Edition, (1977)