Cost-benefit Analysis of a Responsive Sampling Strategy in MEPS

Sadeq R. Chowdhury
Division of Statistical Research and Methods[1]
Agency for Healthcare Research and Quality
5600 Fishers Lane, Rockville, MD 20857

Abstract

Attempting to obtain responses through repeated follow-ups of reluctant respondents both complicates the data collection process and incurs considerable extra costs to the Medical Expenditure Panel Survey (MEPS) Household Component. Due to more extensive follow-up and lower response rates, the costs per completed interview for these households are significantly higher compared to households that respond to initial contacts. A responsive design that subsamples nonrespondents after a reasonable number of follow-ups is being considered as an option to reduce data collection costs in the MEPS. Using survey paradata, this paper presents a cost-benefit analysis of subsampling interim nonrespondents. It discusses potential benefits in terms of costs savings in data collection and increased response rates versus loss in precision of estimates due to increased design effects.

Key words: responsive design, sampling, MEPS, NHIS, subsampling nonrespondents

1. Introduction

The Medical Expenditure Panel Survey (MEPS) has been conducted by the Agency for Healthcare Research and Quality (AHRQ) since 1996. The MEPS provides nationally representative estimates of health care use, expenditures, sources of payment, and health insurance coverage for the U.S. civilian non-institutionalized population. It consists of three survey components with the Household Component (HC) as the core survey. The MEPS Household Component (will be generally referred to as MEPS hereafter) also provides estimates of respondents' health status, demographic and socio-economic characteristics, employment, access to care, and satisfaction with health care. The survey is used to produce estimates for persons and families as well as subgroups of the population. The sample for MEPS is selected from the responding households to the prior year’s National Health Interview Survey (NHIS). A new sample panel for MEPS is selected every year and is followed for two consecutive years and hence two overlapping panels are combined each year to produce annual estimates from a total sample of about 14,000 households and 30,000 individuals. Since the NHIS is based on a multistage area probability sample design and since the MEPS is a subsample of the NHIS, the MEPS sample is also based on a multistage area probability design. The details of the NHIS

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

sample design can be found in Parsons et al. (2014) and the details of the MEPS sample design can be found in Ezzati-Rice et al. (2008).

There are various advantages because of the relationship between the two surveys. The MEPS sampling frame from the NHIS contains a wealth of information collected in the NHIS, including demographic and socio-economic characteristics of responding members. Because of this information on the frame, MEPS does not need to screen households to locate and oversample policy-relevant subgroups of the population. The auxiliary data available on the frame are also used for nonresponse adjustments in MEPS. The linkage or connection of these surveys offers a unique opportunity to use paradata from NHIS to develop appropriate sampling strategies in MEPS. The linked data are also used to expand the analytic capacity of MEPS.

Like many other national household surveys, the response rate in MEPS is decreasing over time and the cost of data collection is becoming more and more expensive due to growing reluctance from households to participate in surveys. To address this, innovative sampling schemes of adoptive/responsive natures (Hansen & Hurwitz, 1946; Groves and Heeringa, 2006) are being explored in MEPS. In recent years, paradata from NHIS have been used to form substrata with differential response propensities and then different sampling rates are used to reduce data collection effort. Previous research has explored disproportionate sampling as a way to lower data collection costs (Barron, et al., 2015). Another responsive sampling approach currently being considered is to subsample the interim nonrespondents or unresolved households. Groves and Heeringa (2006) defined responsive design and extensively discussed the use of paradata to develop responsive designs to control survey costs and nonresponse while maintaining the quality of survey estimates. In recent years paradata are increasingly being used for developing responsive sampling designs (Durrant et al., 2014; Durrant et al. 2015; Kreuter 2013, Wagner, 2013, Groves et al., 2009).

The majority of the respondents to MEPS complete their response within a few initial contacts but the remaining respondents need repeated follow-ups. Also, the response rate of the households that are resolved[2] during initial contacts is much higher than that of the households that require extensive follow-up. Following up these reluctant households not only makes the data collection very expensive but it also complicates the field operations. Due to more extensive follow-up and lower response rates, the costs per completed interview for these households are significantly higher compared to households that respond to initial contacts. Therefore, a responsive design that subsamples interim unresolved or nonrespondents after a reasonable number of follow-ups can be considered as an option to reduce data collection efforts and to increase the response rate in MEPS.

Using paradata from the previous rounds of the survey, this paper presents a cost-benefit analysis of the subsampling scheme discussed above. It discusses potential benefits in terms of costs savings in data collection and increased response rates versus loss in precision of estimates due to increased design effects. The paper compares the cost-benefit of the proposed subsampling scheme with that of the current scheme based on NHIS paradata. The paper also discusses an approach to optimize sample allocation to minimize data collection costs while considering substrata response rates and design effect due to variation in sampling rates.

[2] either as respondent or declared as nonrespondent due to refusal, unlocatable, etc.

2. Responsive Sampling Strategy in MEPS

2.1 General Sampling Scheme

Based on the race and ethnicity information collected in NHIS, minorities are oversampled in MEPS to improve sample sizes for policy-relevant analyses. Minority households are defined as at least one or more people in the household that identify as Hispanic, non-Hispanic Asian, or non-Hispanic black in that hierarchy structure and are sampled with certainty. Non-Hispanic white/other households are the largest sampling domain in MEPS and are sampled at a non-certainty rate that balances the precision requirement of the estimates for this domain and the pre-assigned targeted total sample size for allocation. Table 1 shows the sampling rates used in different domains in recent years from the MEPS frame. Over the past several years, the overall sampling rate for the non-Hispanic white/other households has been about 61% of the households on the frame.

Table 1a: Sampling Rates from MEPS Frame used in Various Sampling Domains in Recent Years

Domain	*Sampling Rate*
Hispanic	100%
Non-Hispanic Asian	100%
Non-Hispanic black	100%
Non-Hispanic white/other	61%

2.2 Current Complete/Partial Stratified Sampling Scheme

In recent years, NHIS paradata have been used to further stratify the non-Hispanic white/other households to help develop a sampling strategy to reduce data collection effort. A good predictor of response propensity in MEPS is the NHIS complete/partial interview status[3]. The households with a complete NHIS interview have much higher response propensity in MEPS than the households with a partial NHIS interview. The NHIS complete/partial interview status is used to form substrata and then differential sampling rates are used to reduce data collection effort. Since minority households are selected with certainty, this strategy is only used for non-certainty households i.e., non-Hispanic white/other households. As Table 1b shows, while the overall sampling rates for Non-Hispanic white/other households is 61%, a higher sampling rate of about 63% is used for the NHIS complete subdomain and a lower sampling rate of about 49% is used for the NHIS partial subdomain in recent years.

Table 1b: Sampling Rates from MEPS Frame used in Subdomains under White/Other Domain in Recent Years

Domain/Subdomain	*Sampling Rate*
Non-Hispanic white/other	61%*
NHIS Complete	63%
NHIS Partial	49%

*This number is a weighted average of the complete and partial sampling rates

[3]In NHIS, a complete interview means the household composition, family, sample adult, and sample child (if a child was present) modules were all completed and a partial interview means that at least a sufficient portion of the family module was completed.

2.3 Subsampling Interim Unresolved Households

A responsive design that subsamples interim unresolved households after a reasonable number of follow-ups is being considered as an option to reduce data collection costs in MEPS. Table 2 shows the distribution of the number of contacts (includes contacts, contact attempts and calls but will be generally referred to as contacts or calls) per household in MEPS panels 17 and 18. While the mean number of contacts for resolution of a household is 8.6, it takes only ≤3 contacts to resolve 25% and ≤6 contacts to resolve 50% of the sampled households. However, after that the number of contacts required for resolution keeps increasing with ≤11 contacts required to resolved 75% households and ≤40 contacts required to resolve 90% households.

Table 2: Distribution of Number of Contacts for Resolution in MEPS Panels 17 and 18

Distribution Parameters	*Number of Contacts/Calls*
Mean	8.59
Percentiles	P25 = 3, P50 = 6, P75 = 11, P90 = 40

Analyzing the distribution of the number of contacts required for resolution, a subsampling scheme is proposed with a cut-off around 10 contacts. All households would be made up to 10 contacts and after 10 contacts only a subsample of the unresolved households would be followed up further and the remaining households would not be followed up any more. The cut-off value of 10 contacts is determined by varying the cut-off and observing the difference in the average number of contacts and the response rate between substrata. A slight change in the cut-off value (say by ±1 or 2) does not make much difference on the effectiveness of the scheme but if the cut-off is reduced too much then the difference in average number of contacts and response rate between substrata will be smaller and the scheme will be less effective. On the other hand, if the cut-off is increased too much then most of the sample will be resolved before reaching the cut-off and the scope for subsampling would be less.

2.3.1 Implementation of subsampling scheme

An important issue is how to implement such a subsampling scheme without disrupting the field operation. One way to do this is by assigning a subsampling flag using the required subsampling rate to all sampled households before going to the field and then all households will be followed-up with up to 10 calls/contacts irrespective of the subsampling flag. After 10 calls if a household is still unresolved, it will be followed-up only if the subsampling flag indicates the household is selected for subsampling. This strategy should avoid any disruption of the field operation that would occur if it were necessary to consult from the field with the sampling statistician.

3. Optimum Allocation or Subsampling

For allocating the sample to substrata formed for developing a cost-effective design, a scheme is described for optimizing sample allocation or optimizing subsampling rates by balancing data collection costs, response rate, and the variance of the estimates. This is done by using a cost function that incorporates a fixed cost and a variable cost of data collection. The variable cost is measured in terms of the number of contacts in each substratum. The average number of contacts in a substratum is used as a proxy for cost in the sample allocation discussion. The number of contacts is affected by many factors, including locating the study participants, willingness of respondents to participate in the survey, and break offs during the survey.

3.1 Cost Function

The data collection cost function for a domain or a broad stratum can be defined as follows:

$$C = C_o + \sum C_h n_h \qquad (1)$$

where C_o is the fixed cost and all other costs that are invariant to subsampling in substratum h, C_h is the average cost for completing each sampled unit in substratum h and n_h is the sample size in substratum h.

The average cost C_h in substratum h can be defined by factoring in the average number of contacts and response rate as follows:

$C_h = Q_h/R_h$ = Average number of contacts for obtaining a response,

with

Q_h= average number of contacts for each selected household including both respondents and nonrespondents,

R_h= $\frac{n_{hr}}{n_h}$ = response rate, where n_{hr} is the number of respondents in substratum h.

Any other perceived or real cost component can be incorporated in deriving C_h or C. For example, any variation in the unit cost of a contact by geography or other factors can also be accounted by computing a weighted average cost C_h.

3.2 Sample Allocation

In the absence of any attempt to reduce the number of contacts, no sampling substratum is formed and there is no need for sample allocation. However, for a comparison with a stratified sampling or subsampling scheme, the sample under no stratification can be considered on expectation as proportionally allocated to whatever strata are formed in an alternative scheme. Therefore, if no substratum is formed for differential sampling or subsampling, then the sample in an overall draw is expected to be allocated proportionally as follows:

$$n_h = n * \frac{N_h}{\sum_h N_h} \qquad (2)$$

where n is the overall sample size in the domain or in a broad stratum, n_h is the expected allocated sample size in substratum h, and N_h is the frame size in substratum h.

On the other hand, to minimize the cost (in terms of the number of contacts) for a fixed sample size n, an appropriate stratification can be formed and the sample can be allocated or subsampled optimally (Neyman, 1934) as follows:

$$n_h = n * \frac{N_h S_h/\sqrt{C_h}}{\sum N_h S_h/\sqrt{C_h}} = n * \frac{N_h S_h/\sqrt{Q_h/R_h}}{\sum N_h S_h/\sqrt{Q_h/R_h}} \qquad (3)$$

where S_h is the standard deviation of a target variable in substratum h .

Since the interest here is to minimize the variance increase due to variation in weights for differential allocation or sampling rates, the variation of a target variable in different substrata within a broad stratum will be assumed the same i.e., $S_h = S$. In that case, the above expression for optimal allocation will reduce to:

$$n_h = n * \frac{N_h/\sqrt{C_h}}{\sum N_h/\sqrt{C_h}} = n * \frac{N_h/\sqrt{Q_h/R_h}}{\sum N_h/\sqrt{Q_h/R_h}} \quad (4)$$

This means that, in consistence with the objective of reducing costs, the allocation is set to sample more heavily within substratum that have larger populations and lower costs (Lohr, 2009). Under a subsampling scheme, an appropriate sample size will be selected so that the expected number of households resolved within 10 contacts is equal to the optimum number as determined by the above optimum allocation formula. Then the households not yet resolved after 10 contacts will be subsampled at a rate so that the number subsampled is equal to the allocated number for the more than 10 contacts substratum. That means the optimum allocation formula will determine the optimum subsampling rate.

The above allocation will minimize costs for a fixed sample size n in a domain/broad stratum. However, as the sampling rate varies by substrata the variance in the stratum will increase due to variation in weights. So to keep the variance fixed, the stratum sample size should be adjusted by considering the higher design effect and increase in response rate.

3.3 Coefficient of Variation of Sampling Weights

As it deviates from the proportional allocation to the optimum allocation to minimize costs, the variation in base sampling weights (w) will increase the overall design effect (*deff*) as follows (Kish, 1965):

$$deff = (1 + CV_w^2) \quad (5)$$

where $CV_w = \frac{\sqrt{V(w)}}{\bar{w}}$ is the coefficient of variation of sampling weights across substrata with the variance of weight,

$$V(w) = \sqrt{\frac{\sum_h n_h (w_h - \bar{w})^2}{n}} \quad (6)$$

For the proportional allocation, since the sampling rate is the same in all substrata, the CV_w=0 and hence *deff*=1; the effective sample size will remain the same as n, where n is the nominal (realized) sample size. On the other hand, under the optimum allocation, the effective sample size will be reduced to $n/deff$.

This loss in the effective sample size will be considered as the cost in the cost-benefit analysis of the proportional and the optimum allocation.

3.4 Adjusting Stratum Sample Size to Control Variance

Considering the increased design effect and increased response rate, the stratum sample size n can be adjusted as follows:

$$n^* = n \, \frac{R\delta}{R^*} \quad (7)$$

where n^* is the adjusted sample size, R is the stratum-level response rate with equal sampling rate across the stratum, R^* is the stratum-level response rate under the above allocation and δ is the design effect for variation in sampling rate by substrata. The adjusted sample size n^* can now be used in (4) and reallocate the adjusted sample size.

For appropriate stratification and optimal allocation, the ratio of increase in stratum-level response rate $\left(\frac{R}{R^*}\right)$ is usually higher than the increase in design effect (δ) i.e., $\frac{R}{R^*} \leq \partial$, implying $n^* \leq n$. That means if the response rate increases more than the increase in the design effect then the effective responding sample size will increase and the stratum sample size can be reduced while maintaining the same level precision of the estimates. Therefore, an optimal allocation with appropriate stratification can reduce data collection costs and also increase response rate while keeping the stratum-level variance fixed.

4. Evaluation of Alternative Schemes

The evaluation of the overall benefits is made in terms of lower data collection costs and improved response rate and the expected inflation in variance due to the increase in the CV of the weights. For evaluation, the cost-effectiveness for comparing alternative schemes will be simulated using combined data from MEPS Panels 17 and 18, which will serve as the sampling frame, for which the number of contacts and response status for households are known.

Each of the two sampling schemes i.e., the current scheme with NHIS complete/partial stratification and the subsampling of interim unresolved households will be compared with the default scheme in which no responsive sampling effort is made. The performances of the two schemes will then be evaluated by comparing their performances against the default scheme.

Since the sampling under no stratification is on expectation equal to a proportional allocation, the comparison of the cost-effectiveness of a non-stratified scheme and that of a stratified scheme with optimal allocation comes down to the comparison between a proportional and an optimum allocation.

For evaluating the currently used complete/partial scheme, the cost-effectiveness will be simulated using a hypothetical sample of size 4,750 households, equal to the usual sample size selected for non-certainty households in MEPS, because the current scheme is applied to the non-certainty domain only. The sample is assumed to be selected by allocating the total sample across the complete/partial substratum using the optimum allocation procedure described above.

For evaluating the proposed subsampling scheme after 10 contacts, a sample of size 9,700 households will be hypothetically selected. This sample size is roughly equal to the usual sample size selected from all domains in MEPS, because the scheme is applicable to all domains. For implementing this scheme, the initial sample size must be inflated so that a subsampling can be applied later. First, the total sample size will be inflated so that the optimum number of households can be resolved within 10 contacts as determined by the optimum allocation and then an optimum subsampling rate can be applied so that only an optimum number of households is followed-up after 10 contacts. The total sample size after subsampling will be equal to the original target of 9,700.

First, the evaluation will be presented for the currently used complete/partial scheme and then the evaluation will be presented for the subsampling scheme of interim unresolved households. The response rates and the mean number of contacts observed in Panels 17 and 18 will be used for the evaluation.

5. Evaluation Results

5.1 Complete/Partial Stratification Scheme

Table 3 shows the distribution of the sample, response rates and mean number of contacts by NHIS interview status for the non-certainty sampling domain of MEPS Panels 17 and 18 combined. The households with completed NHIS interviews have a higher response propensity and a lower number of average contacts per complete in MEPS. The MEPS response rate is 76.4% for NHIS completes compared to 58.5% for NHIS partials. The average number of contacts per response in MEPS is much lower (10.4) among NHIS completes compared to the average number of contacts (17.2) for NHIS partials.

Table 3: Sample size, Response Rate, Total and Average Contacts in MEPS by NHIS Complete/Partial Interview Status

NHIS Outcome	*Number of Households*		*Response Rate*	*Total Contacts*	*Average contacts per household*	*Average contacts per complete*
	Sampled	*Responded*				
Complete	6,599	5,042	76.4%	52,335	7.9	10.4
Partial	1,183	692	58.5%	8,788	10.0	17.2
Total	7,782	5,734	73.7%	64,224	8.3	11.2

Table 4 presents a comparison of sample allocation and the cost-benefit evaluation criteria between the current complete/partial stratification scheme with optimum allocation and the default no-stratification scheme. Under the default scheme, there is no sub-stratification but the expected numbers based on a proportional allocation are presented for comparison with the subsampling scheme. The sampling rates are equal from both the NHIS complete and partial strata (61.0%) while under optimal allocation the sampling rates are 63.2% and 49.2% respectively. This difference in sampling rates under the optimum allocation is due to the higher cost in terms of the number of contacts in the partial stratum, which drives the sample allocation to be lower in the partial stratum and higher in the complete substratum.

As a result, the total number of contacts is expected to go down from 39,201 under the default scheme to 38,906 under the complete/partial scheme. Similarly, since the response rate is lower in the partial substratum, the overall response rate is expected to be slightly higher under the optimal allocation (74.2%, 3,525 respondents) than under the default no-stratification (73.7%, 3,500 respondents). However, due to the 9% increase in variation in weights under the complete/partial scheme, the effective responding sample size will come down to 3,497. On the other hand, since there is no additional variation in weights under the no-stratification scheme because the sampling rate is the same in both sampling substrata, the effective sample size will remain the same at 3,500. Since the effective responding sample size under both schemes are almost the same (3,500 and 3,497), no adjustment (as discussed in section 3.4) to the overall sample size is made under the stratification scheme. Therefore, while the effective sample size remains almost the same under both methods, the total number of contacts under the stratification scheme is lower and the response rate is slightly higher.

Table 4: Comparison of Current Stratification Scheme with Default Scheme

NHIS Outcome	*Default: No Stratification*				*Stratification with Optimum Allocation*			
	Sampled	*Sampling Rate*	*Response*	*Total Contacts*	*Sampled*	*Sampling Rate*	*Response*	*Total Contacts*
Complete	4,028	61.0%	3,078	31,957	4,169	63.2%	3,185	33,072
Partial	722	61.0%	422	7,245	581	49.2%	340	5,834
Total	4,750	61.0%	3,500	39,201	4,750	61.0%	3,525	38,906
Response Rate			73.7%				74.2%	
CV of weights			0%				9.0%	
Effective Sample Size*			3,500				3,497	
Difference in Response Rates							+0.50% pt.	
Difference in Total Contacts							-295 (-0.75%)	
Difference in Effective Responding Sample Sizes							-3	

*Effective sample size = Total number of respondents/(1+CV^2) = 3,525/(1+(0.09^2)

5.2 Subsampling Scheme

Table 5 shows the mean number of contacts and the response rate for households resolved within or after 10 contacts. For households resolved (i.e., classified as definite respondent or nonrespondent) within 10 contacts, the mean number of contacts per household is lower (6.0), response rate is much higher (87.4%) and as a result the mean number of contacts per complete (9.6) is lower than for households resolved after 10 contacts for which the corresponding numbers are 9.3, 48.5%, and 19.1 respectively.

Table 5: Sample Size, Response Rate, Total and Average Contacts by Subsampling Strata Formed Using ≤10 and >10 Contacts in MEPS Panels 17 & 18 Combined

Substrata	*Number of Households*		*Response*	*Total*	*Mean contacts*	*Mean contacts*
	Resolved	*Response*	*Rate*	*Contacts*	*per household*	*per complete*
≤10 Contacts	14,013	12,249	87.4%	116,938	6.0	9.6
>10 Contacts	5,339	2587	48.5%	49,391	9.3	19.1
Total	19,352	14,836	76.7%	166,329	8.6	11.2

Table 6 shows how the sample would be selected and subsampled under the subsampling scheme. Of the 19,352 households on the frame, 14,013 households are expected to be resolved within 10 contacts and the remaining 5,339 are expected to be resolved after 10 contacts. Under the optimum allocation, roughly 7,641 households must be resolved within 10 contacts and 2,059 must be resolved after 10 contacts for the total targeted sample size of 9,700 resolved households. Since the target is to achieve 9,700 resolved households after subsampling, the target sample size must be inflated first to allow subsampling later. Secondly, the sample size must be inflated to such an extent so that the optimum allocation target of 7,641 households can be resolved within 10 contacts. Therefore, the target sample size of 9,700 resolved households is first inflated to 10,553 so that 7,641 households can be resolved within 10 contacts and the remaining 2,911 households are subsampled at a rate of 70.7% to select 2,059 households for further follow-up to achieve the final total sample size of 9,700 resolved households.

Table 6: Sample Selection with Optimum Allocation under Subsampling Scheme

Substrata	*Frame*	*Optimum Sample Allocation*	*Optimum Sampling Rate*	*Subsampling*			
				Inflated Sample Size	*Sub-sampling Rate*	*Sub-sample Size*	*Final Sampling Rate*
≤ 10 Contacts	14,013	7,641	54.5%	7,641	100%	7,641	54.5%
>10 Contacts	5,339	2,059	38.6%	2,911	70.7%	2,059	38.6%
Total	19,352	9,700	50.1%	10,553	-	9,700	50.1%

Tables 7 and 8 present the sample allocation and a cost-benefit analysis of the subsampling of interim unresolved households in comparison to the default no stratification/subsampling scheme. Table 7 shows that if the sample size is 9,700 resolved households then under the default scheme, 7,437 households are expected to respond based on the overall expected response rate of 76.7%. There will be no sub-stratification under this scheme but the expected numbers based on a proportional allocation are presented just for comparison with the subsampling scheme. The table also shows that under the subsampling scheme, based on the substratum response rates (87.4% and 48.5%) as observed in MEPS panels 17 and 18, there will be 6,679 respondents in the ≤10 contacts substratum and 998 respondents in the >10 contacts substratum resulting in a total response from 7,677 households that corresponds to an overall response rate of 79.1%. That means, under the subsampling scheme there will be 240 more responding households which corresponds to a 2.4% point higher response rate compared to the default scheme. However, this increase in response rate comes with a cost in terms of increase in the design effect. The design effect is 1.024 under the subsampling scheme due to the variation in the sampling rate by substrata compared to the design effect of 1.0 under the default scheme which has no variation of sampling rate by substrata. If the responding sample size is adjusted for the design effect, the effective responding sample size for the default scheme will remain the same at 7,437 but it will decrease to 7,495 under the subsampling scheme, still a difference of +59 under the subsampling scheme that resulted mainly due to the higher response rate.

Table 7: Comparison of Subsampling Scheme with Default Scheme before Adjusting Effective Responding Sample Size

Implicit Substrata	*Default (No subsampling)*			*Subsampling*		
	Sampled	*Response Rate*	*Response*	*Optimum Allocation*	*Response Rate*	*Response*
≤ 10 Contacts	7,024	87.4%	6,140	7,641	87.4%	6,679
>10 Contacts	2,676	48.5%	1,297	2,059	48.5%	998
Total	9,700	76.7%	7,437	9,700	79.1%	7,677
Difference in Response Rate					+2.4% pt.	+240
Design Effect			1.00			1.024
Effective Responding Sample Size			7,437			7,495 (+59)

Table 8 presents the analysis again by adjusting the sample size under the subsampling scheme so that the effective responding sample sizes under both schemes are the same to allow a fair cost (number of contacts) comparison. The sample size for the subsampling scheme has been reduced by 59 to 9,641resulting an effective sample size of 7,437 under both schemes. A comparison of the total numbers of contacts under the two schemes with the same effective responding sample size (i.e., fixed variance) shows that total number of contacts required under the subsampling scheme is 1,106 (1.3%) lower than that under the default scheme. The response rate is also higher by 2.6% point under the subsampling scheme.

Table 8: Comparison of Subsampling Scheme with Default Scheme after Adjusting Effective Responding Sample Size

Implicit substrata	*Default (No subsampling)*				*Subsampling*			
	Sample	*Response Rate*	*Response*	*Contacts*	*Sample*	*Response Rate*	*Response*	*Contacts*
≤10 Contacts	7,024	87.4%	6,140	58,614	7,641	87.4%	6,679	63,766
>10 Contacts	2,676	48.5%	1,297	24,757	2,000	48.5%	969	18,499
Total	9,700	76.7%	7,437	83,371	9,641	79.3%	7,648	82,265
Design effect			1.00				1.029	
Effective Responding Sample Size			7,437				7,437	
Difference in Effective Responding Sample size							0	
Difference in Response Rate							+2.6% pt.	
Difference in Total Contacts							-1,106 (-1.3%)	

Table 9 summarizes the comparison results for the current complete/partial stratification scheme and the subsampling scheme in relation to the default no stratification scheme. It shows that the response rate can be increased by 2.6% point and the total number of contacts can be reduced by 1.3% under the subsampling scheme. These rates are better than the response rate increase of 0.5% point and the decrease in the number of contacts of 0.75% under the current scheme. However, although the difference of 2.1% points in the response rate is noticeable, the difference of 0.55% point in the number of contacts is not appreciable under the subsampling scheme.

Table 9: Summary of Comparisons of Current Stratification Scheme and Subsampling Scheme with Default Scheme

Evaluation Criterion	*Complete/ Partial Stratification (Current) [1]*	*Subsampling after 10 contacts (Proposed) [2]*	*Difference [2]-[1]*
Difference in Response Rate (% point)	+0.5%	+2.6%	+2.1
Difference in Total Contacts	-0.75%	-1.3%	-.55%

The reduction in the number of contacts is not as large as expected under the subsampling scheme because the strategy requires at least 10 contacts to all unresolved households before subsampling. The costs of 10 contacts to the households that are subsampled out and not followed any further increases the overall cost per complete of the subsampled households, because the average number of contacts per complete is the sum of the total number of contacts made to both resolved and unresolved households divided by the number of completes. So it appears that if making these 10 contacts could somehow be avoided before starting the subsampling, the gain in terms of reducing the number of contacts could be much higher. One way of doing this would be if the households that have a lower propensity of being resolved during initial calls could be identified earlier and subsampled at a lower rate at the time of sampling i.e., before going to the field. In other words, pre-stratifying and sampling at different rates as done under the current scheme would be more effective than subsampling if the stratification can be done more effectively by using information other than just complete/partial interview status. One idea is to utilize all available background characteristics and information from the linked NHIS and develop a model to predict response propensities of the households during the initial (say 10) contacts. Then using those predicted propensities, substrata can be formed by separating the households with high and low propensity of cooperation within initial calls. To give an idea of the potential savings for using such a scheme, Table 10 shows the savings that would be achieved if the substrata (≤10, >10 contacts) could be formed before sampling and households selected at optimal rates. It shows that while the response rate would remain at the same level, the savings in the number of contacts can be up to 12%. Therefore, if the response propensity substrata can be formed using available information before hand and the sampling is done at differential rates based on optimal allocation, the possibility of savings could be much higher compared to a responsive design subsampling scheme.

Table 10: Comparison of Prestratified (based on Predicted Propensity) Sampling Scheme with Subsampling Scheme

Evaluation Criterion	*Subsampling after 10 contacts*	*Prestratified Sampling*
Difference in Response Rate (% point)	+2.6%	+2.6%
Difference in Total Contacts	-1.33%	-12.3%

6. Conclusion

A cost-benefit analysis of a responsive sampling scheme for MEPS where interim unresolved households are subsampled to reduce data collection effort is presented. A method for implementing the subsampling scheme without disrupting the field operations is discussed. Also a method for determining optimal sample allocation by considering variation in response rates and costs of data collection in different substrata is presented. An approach to evaluate cost-effectiveness in terms of design effect, data collection effort and response rates for different alternative schemes using paradata from previous rounds of the survey is also presented. The effectiveness of the proposed subsampling scheme is compared with the current scheme where substrata are formed using NHIS complete/partial interview status.

The evaluation shows that, with an optimum subsampling rate, the subsampling of unresolved households after some initial contacts can increase the unweighted response rate and can decrease data collection costs while maintaining the effective responding sample size. It also has the potential to increase the weighted response rate as the collection effort will be concentrated on a smaller sample of hard to reach subgroup. The comparison of the subsampling scheme with the current scheme shows that the subsampling scheme can increase the response rate and reduce collection costs more than the current scheme. However, although the increase in the response rate is appreciable, the decrease in data collection costs is not.

Under the subsampling scheme, the reduction in the number of calls is not significant as the scheme requires making a certain number of calls to all households before identifying the substratum for subsampling. The savings in the number of calls could be much higher if the households that would need higher calls could be identified earlier and sampled at an optimally lower rate than subsampling later. In other words, there is a potential for higher savings if the current scheme of pre-stratifying and selecting at different rates can be improved by using multiple variables in the pre-stratification. Further research is planned to explore this idea. For example, using all available paradata and other characteristics of the households, it may be possible to develop effective models for response propensity and for predicting the number of calls required for resolution. The predicted number of calls or propensity scores can be used for pre-stratification and selecting at different rates.

References

Barron, M., Davern, M., Montgomery, R., Tao, X., Wolter, K. M., Zeng, W., . . . Black, C. 2015. Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations. *Journal of Official Statistics, 31*(4), 545–557.

Durrant, G. B., Maslovskaya, O., and Smith, P. W. F. 2014. "Sequence Analysis as a Tool for Investigating Call Record Data," Working paper, University of Southampton.

Durrant, G.B., Maslovskaya, O., Smith, P.W. 2015. Modelling Final Outcome and Length of Call to Improve Efficiency in Call Scheduling, *Journal of Survey Statistics and Methodology*, forthcoming.

Ezzati-Rice, T.M., Rohde, F., Greenblatt, J. 2008. *Sample Design of the Medical Expenditure Panel Survey Household Component, 1998–2007*. Methodology Report No. 22. March 2008. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr22/mr22.pdf

Groves, R.M and Heeringa, S.G. 2006. Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society,* Series A, **169**, *Part* 3, *pp.* 439–457

Groves, R.M., Mosher W.D., Lepkowski J, Kirgis N.G., 2009. Planning and development of the continuous National Survey of Family Growth. National Center for Health Statistics. Vital Health Stat, **1(48)**. 2009.

Hansen, M. H., & Hurwitz, W. N. 1946. The Problem of Non-Response in Sample Surveys. *Journal of the American Statistical Association,* ***41*****(236)**, 517-529.

Kish, L. 1965. *Survey Sampling.* New York: Wiley.

Kreuter, F. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information,* edited by Frauke Kreuter, John Wiley & Sons, Inc, Hoboken, NJ.

Lohr, S. L. 2009. *Sampling: Design and Analysis.* Boston: Richard Stratton.

Neyman, J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558-606.

Parsons, V.L., Moriarity, C., Jonas, K., Moore, T.F., Davis, K.E., Tompkins, L. 2014. Design and estimation for the National Health Interview Survey, 2006–2015. National Center for Health Statistics. *Vital Health Stat* 2(165). 2014.

Wagner, J. 2013. Using Paradata-Driven Models to Improve Contact Rates in Telephone and Face-to-Face Surveys, in *Improving Surveys with Paradata: Analytic Use of Process Information*, ed. Kreuter, F., New Jersey: John Wiley and Sons, pp. 145-170