SAMPLING AND ESTIMATION ISSUES IN THE MEDICARE CURRENT BENEFICIARY SURVEY

Julie O'Connell, Annie Lo, David Ferraro, R. Clifton Bailey
Julie O'Connell, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Composite estimation, multistage area probability sampling, longitudinal studies, Medicare population, rotating panels.

1. Introduction

Disclaimer: The opinions are those of the authors and do not represent policies or opinions of HCFA.

1.1 MCBS Sample Design

The MCBS is a continuous, multi-purpose panel survey of Medicare beneficiaries that is intended to provide data on health care use and costs. The survey is sponsored by the Health Care Financing Administration (HCFA). The sample design for the MCBS is a stratified area probability design with three stages of selection.

Primary sampling units (PSUs) consist of MSAs or clusters of non-metropolitan counties in the 50 states, the District of Columbia and Puerto Rico. Within region and metropolitan status, the PSUs were grouped into sampling strata. The strata were constructed to be internally homogeneous with respect to socioeconomic characteristics and of roughly equal population size. Large metropolitan areas were sampled with certainty, resulting in 33 "certainty" PSUs. From each of the remaining 37 "noncertainty" strata, two PSUs were sampled, for a total of 107 PSUs.

The second sampling stage consists of ZIP code areas within each sampled PSU. In order to simplify linking with county-level data, sampling units for the second stage consist of ZIP codes areas within a single county. ZIP code areas that cross county borders are split by county into separate units or ZIP fragments. These ZIP fragments are then combined into clusters for sampling, so that a reasonable aggregate measure of size is achieved for each cluster. Clusters consist of ZIP fragments within a stratum that are similar with respect to socioeconomic characteristics. Each year, the set of sampled ZIP fragments is supplemented to include newly created ZIP code areas.

The primary ZIP cluster sample selected in 1991 consisted of 4,423 sampled ZIP fragments in 1,163 clusters. The 1992 sample included ZIP clusters sampled for coverage improvement as well as newly created ZIP codes. Through 1997, 312 ZIP clusters have been added to the sample including 929 ZIP fragments for a total of 1,475 ZIP clusters including 5,352 ZIP fragments.

At the third sampling stage, Medicare beneficiaries are sampled within each sampled ZIP cluster. The beneficiary sample is stratified within seven age categories: 0-44, 45-64, 65-69, 70-74, 75-79, 80-84, 85+. The target sample size for the continuing annual sample is 12,000 responding beneficiaries, including 1,000 beneficiaries in each of the disability age categories and 2,000 beneficiaries in each of the remaining categories. Young disabled (0-44) and very old (80-84, 85+) are oversampled; age 65-69, 70-74, 75-79 categories are undersampled.

2. MCBS Samples

An initial MCBS sample, consisting of 15,411 beneficiaries was selected in 1991. First MCBS interviews were conducted in the fall of 1991. In each of the following two years, supplementary samples were selected to include newly eligible beneficiaries and to maintain sample size in each age stratum. These supplements consisted of 2,410 and 2,449 sampled beneficiaries, respectively. First interviews for each supplement are conducted in the fall of the year in which the supplement is selected.

MCBS beneficiaries are interviewed three times a year, and each interview round is administered over a 4-month period. After two years of interviews, it became apparent that the enormous respondent burden imposed by an unending sequence of interviews was adversely affecting the cumulative response rate. Although the initial rate of response for the 1991 sample was 87 percent, by Round 8 the cumulative response rate had dropped to 70 percent and by Round 12 the cumulative response rate for this sample was 65 percent.

In 1994, it was decided to move to a rotating panel design in which each annual supplement is selected as a nationally representative sample. Under this design sampled beneficiaries remain in the sample for four years and are then released. In order to maintain sample sizes in the continuing sample, approximately 6,000 beneficiaries are needed in each annual supplement.

Procedures to convert the MCBS sample to a rotating panel design were initiated with the 1994 supplement. For 1994 through 1997, each annual supplement was selected as a nationally representative sample. Sample sizes for these supplements ranged from 6,349 beneficiaries in 1994 to 6,599 beneficiaries in 1997. Beneficiaries in the 1991, 1992 and 1993 supplements were phased out of the sample over two years from Round 13 (fall, 1995) to Round 19 (fall, 1997). Approximately one-third of the beneficiaries in these panels were released each year. Round 19 was the first interview in which the rotating panel design was fully in effect.

The most recent supplements have included special one-time augmentations for HCFA's Office of Research and Demonstrations (ORD) analyses of Medicare HMOs. These ORD/HMO augmental supplements are interviewed only once at the fall interview round, and the data are included in the access to care files. The augmentation consists of additional sampled cases in selected target areas, as well as additional beneficiaries in risk HMO plans nationally. The supplements for 1996 and 1997 both included an ORD/HMO augmentation. For 1996 the target areas were South Florida and Southern California; for 1997 the target areas were Philadelphia and Phoenix. An augmentation is also planned for 1998 with target areas of Denver, Minneapolis and South Florida.

3. Sampling/Weighting Considerations

Under the rotating panel design, each annual MCBS supplement is selected as a nationally representative sample that represents the population of beneficiaries who are alive and eligible as of January 1 of the current year. Initial interviews for each new supplement are conducted in the fall. Complete annual data are provided for three years, starting with the year following supplement selection.

Each year, HCFA assembles two data files containing MCBS data as well as information from HCFA's administrative database. "Cost and use" files are intended primarily for estimates of charges and payments for a complete calendar year; whereas "access to care" files focus on data that describes access to and satisfaction with health care services. Samples for these data files are comprised of sampled beneficiaries from several different MCBS panels. Weighting adjustments for each sample include adjustments to account for overlap in the panels, so that weighted totals will represent the appropriate population. This section discusses some issues related to selecting and assembling MCBS samples.

3.1 Cost and Use Samples

Target populations for MCBS "cost and use" estimates include beneficiaries in the "ever enrolled" population for a particular calendar year. This population includes beneficiaries who are newly eligible during the calendar year, as well as eligible beneficiaries who die and beneficiaries who are continuously eligible throughout the time period.

Under the rotating panel design, MCBS panels for years *t-1*, *t-2*, and *t-3* will have complete survey data for year *t*. These panels will represent beneficiaries alive and eligible as of January 1 of years *t-1*, *t-2*, and *t-3*, respectively. With appropriate weighting adjustments and subsampling, these samples can be combined to represent the population of beneficiaries who are alive and eligible as of January 1 of year *t*. This combined sample is referred to as the year *t* "continuing sample".

In order to have complete representation for cost and use for year *t*, the sample needs to include representation of beneficiaries who became eligible during year *t-1* and had some eligibility for year *t* and beneficiaries who became eligible during year *t*. Sampled beneficiaries representing these subpopulations are obtained from panels sampled in years *t* and *t+1*, respectively. Since they entered the MCBS in the fall of year *t* or year *t+1*, these sampled beneficiaries have no survey data on charges and payments for year *t*, so that year *t* charge and payment data must be imputed. These sampled beneficiaries are referred to as year *t* "ghosts".

Year *t* deaths of beneficiaries who became eligible on or before January 1 of year *t-1* are represented by year *t* deaths in the continuing sample. Year *t* deaths of beneficiaries who became eligible during year *t-1* are represented by year *t* deaths of "ghosts" in the year *t* panel. In order for the cost and use sample to have representation of year *t* deaths of beneficiaries who became eligible during year *t*, these deaths must be included in the panel for year *t+1*.

Thus each supplement is augmented to include newly eligibles during the previous year regardless of vital status on January 1 of the current year. This also means that records for beneficiaries who die during the previous calendar year must be retained in the sampling frame so that these beneficiaries have a chance of being selected.

3.2 Access to Care Samples

The target population for the "access to care" estimates is the "always enrolled" population for a particular calendar year. This population includes beneficiaries who are enrolled throughout the entire year but excludes beneficiaries who die or lose eligibility during the year and beneficiaries who become eligible after January 1.

Under the rotating panel design, the "access to care" sample for year *t* is comprised of beneficiaries from year *t-3*, *t-2*, *t-1* and *t* panels. Each "access to care" data file includes sampled beneficiaries in these panels who are alive and eligible for the fall interview and who complete this interview. Weighting adjustments include adjustments to account for the overlap in the samples. A flag is added to the file to indicate vital status as of December 31. Weighted estimates from the sample, subsetted to exclude deaths prior to December 31, will reflect the "always enrolled" population for the relevant calendar year.

The panels for years *t-3*, *t-2*, *t-1* can be combined and subsetted to represent that population of beneficiaries who became eligible on or before January 1 of year *t-1*, who remained alive and eligible throughout year *t*. Since sampled beneficiaries in this sample have complete survey data starting with fall of year *t-1*, this sample can be used for longitudinal analyses that cover the one-year period from fall of year *t-1* to fall of year *t*. This is the "one-year backward longitudinal" sample for year *t*.

"Backward longitudinal" samples for year *t* are also available for the two-year period from fall of year *t-2* through fall of year *t* and for the three-year period from the fall of year *t-3* through fall of year *t*. These samples contain beneficiaries from the year *t-2* and *t-3* supplements and represent slightly different populations. Table 1 presents some information on the "backward longitudinal" samples that are available in the access to care releases.

4. Composite Estimates for MCBS Cost and Use Data

Composite estimation is a technique that incorporates information from previous time periods into estimates for the current time period. Each new estimate is computed as a weighted average of two estimates. One estimate is the sample estimate for the current time period; and the second estimate is based on the previous time period and the year-to-year change. Thus,

$$\hat{x}_t = (1-K)x_t'' + K(\hat{x}_{t-1} + d_{t,t-1}) \qquad (1)$$

where

$\hat{x}_t$ = composite estimate for the current time period;

x_t'' = sample estimate for the current time period;

$\hat{x}_{t-1}$ = composite estimate for the preceding time period;

$d_{t,t-1}$ = difference between current and preceding time periods based on units common to both samples = $x_t''' - x_{t-1}'''$;

x_t''' = the estimate for the current time period for units common to both samples;

x_{t-1}''' = estimate for the preceding time period for units common to both samples;

and K is a factor between 0 and 1.[1]

Where there are exactly *N* prior time periods of data available, the estimate becomes a weighted sum of the *N+1* estimates and *N* differences:

$$\begin{aligned}\hat{x}_t &= (1-K)x_t'' + K(d_{t,t-1} + \hat{x}_{t-1}) \\ &= (1-K)x_t'' + K(d_{t,t-1} + (1-K)x_{t-1}'' + K(d_{t-1,t-2} + \hat{x}_{t-1})) \\ &\vdots \\ &= \sum_{i=0}^{N-1}\left[K^i(1-K)x_{t-i}'' + K^{i+1}d_{t-i,t-i-1} + K^N x_{t-N}''\right] \qquad (2)\end{aligned}$$

where x_{t-i}'' and $d_{t-i,t-i-1}$ are the sample estimate and change from the preceding time period for the *t-i*-th time period, *i=0,...,N-1*. For the estimate at *t-N*, no composite estimate is available, so that $\hat{x}_{t-N} = x_{t-N}''$.

Predictions from analytical models suggest that composite estimation will improve precision when year-to-year correlation is high, and improvements will be greater for estimates of differences than for estimates of totals.

Figures 1 and 2 show predictions of variance improvements for estimates of totals and estimates of change, respectively. Values plotted are the ratio of the variance for the composite estimate to the variance for the simple estimate. For this model, variance is assumed to be proportional to sample size. Composite estimation has been used in estimates for the Current Population Survey, with results that are generally consistent with our model predictions.

Amount of precision improvement also depends on relative sizes of the overlap and non-overlap samples and on the variances of the estimates in each subsample, so that it is difficult to predict before hand when composite estimation is most effective.

4.1 MCBS Overlap Samples

We would like to apply the composite estimation technique to estimates from MCBS cost and use samples. To do this we need to identify the overlap in the cost and use samples for years *t* and *t-1*.

[1] *The Current Population Survey: Design and Methodology*. Technical Paper 40, January 1978, Department of Commerce, Bureau of the Census.

In addition, we need to identify an appropriate set of weights for the overlap sample, since the weighting adjustments for nonresponse and combination of panels usually result in weights for a sampled beneficiary that are different each year.

It turns out that the overlap sample for cost and use in years *t* and *t-1* is essentially the two-year "backward longitudinal" sample from the year *t* access to care file. Sampled beneficiaries in this two-year "backward longitudinal" sample will have cost and use data that covers both of these years, so that we can use this sample to estimate year *t-1* to year *t* change. Two-year "backward longitudinal" weights are appropriate for making estimates from this sample; however, the population represented is beneficiaries continuously eligible during years *t* and *t-1*. Thus estimates of change based on this sample are estimates of net change that reflect only the continuously eligible portion of each "ever eligible" population.

We can obtain the difference estimates we need by augmenting each "backward longitudinal" sample to include representation of deaths and newly eligible beneficiaries each year. Sampled beneficiaries for the augmentation can be obtained from each cost and use file. Differences in estimates from the resulting augmented "backward longitudinal" files will be estimates of net change from year *t-1* to year *t*, and these estimates can be used with cross-sectional estimates from each cost and use file to form composite estimates.

We will need to create two separate augmented "backward longitudinal" samples for each difference estimate. The first augmented "backward longitudinal" sample will consist of the original two-year "backward longitudinal" sample for year *t* plus beneficiaries in the cost and use file for year *t* who died, lost entitlement or became eligible during year *t*. This sample will represent the "ever enrolled" population for year *t*. Beneficiaries in this sample will have cost and use data for year *t*.

The second augmented file will consist of the original two-year "backward longitudinal" sample for year *t* plus beneficiaries in the year *t-1* cost and use file who died, lost eligibility or became eligible during year *t-1* and beneficiaries in the year *t-1* cost and use file who died or lost eligibility during year *t*. This sample will represent the "ever enrolled" population for year *t-1*. Beneficiaries in this sample will have cost and use data for year *t-1*.

The modified difference estimates are computed from

$$d^{*}_{t,t-1}{}^{(bl^{*})} = x_t^{\prime\prime\prime\,(bl+ne_t+d_t)} - x_{t-1}^{\prime\prime\prime\,(bl+ne_{t-1}+d_{t-1}+d'_t)} \qquad (3)$$

where $x_t^{\prime\prime\prime\,(bl+ne_t+d_t)}$ is the estimate from the "backward longitudinal" sample with the first augmentation, and $x_{t-1}^{\prime\prime\prime\,(bl+ne_{t-1}+d_{t-1}+d'_t)}$ is the estimate for the "backward longitudinal" sample with the second augmentation.

The resulting composite estimate using data from one prior year is

$$\hat{x}_t = (1-K)x_t'' + K\left(d^{*}_{t,t-1} + x''_{t-1}\right) \qquad (4)$$

Using two-year "backward longitudinal" samples from years *t-i*, *i=0,...,N-1,* with appropriate augmentations, this expression can be expanded to incorporate *N* prior years of data. The form of the resulting composite estimate is similar to (2).

$$\hat{x}_t = \sum_{i=0}^{N-1}\left[K^i(1-K)x''_{t-i} + K^{i+1}d^{*}_{t-i,t-i-1}\right] + K^N x''_{t-N} \, (5)$$

The estimates $d^{*}_{t-i,t-i-1}$ are produced using two-year "backward longitudinal" samples from year *t-i*, appropriately augmented with deaths and newly eligible from years *t-i* and *t-i-1*. The estimates x''_{t-i} are produced using the full sample from each cost and use file, so that no modifications are required for these estimates. Estimates that are produced by this method are appropriate for the population of "ever enrolled" beneficiaries each year.

4.2 Composite Estimate Results

We applied composite estimation to compute estimates for CY 1995 using data from MCBS files for 1993, 1994, and 1995. Estimates included demographic characteristics and health status as well as total 1995 charges and payments by type. Since we expect that improvements in precision will be most apparent for group estimates for which available sample sizes are small, we computed estimates for four analysis subdomains: beneficiaries in risk-HMO plans, beneficiaries residing in nonmetropolitan areas, beneficiaries 85+ years of age, and beneficiaries with dual Medicare-Medicaid eligibility.

Table 2 shows relative variance efficiency for selected estimates. Values of relative efficiency shown in the table are the ratio of variance estimates for the composite estimate to the variance of the simple estimate based on only the 1995 "cost and use" sample. Thus, values less than one indicate variance improvements with compositing, whereas values greater than one indicate loss of precision.

5. Discussion

We expect that improvements in precision will be most evident where year-to-year correlation is high, and we also expect to see more improvement for estimates of year-to-year change than for estimates of CY totals. Our results are generally consistent with these expectations.

Precision improvements for CY estimates were consistently apparent only for education and marital status, both of which have very high year-to-year correlations. Compositing appeared to be least effective for charge and payment estimates, which have more year-to-year variability. Estimates in the table are for *K*=0.5; however, we did not see much difference using an optimal value of *K*. Also estimates in the table are based on a one-year composite. We computed composite estimates with two prior years of data; however, results were similar to those of the one-year composites.

Finally, we note that the effects of imputation on composite estimates are complex and difficult to identify. This is particularly true for data on charges and payments, since these data are all imputed for the "ghosts" in each "cost and use" sample. In general, imputation tends to decrease estimates of variance," so that we would expect this effect both in the simple estimates and the composite estimates; however, the effect on relative efficiency is difficult to assess.

6. References

Apodaca, R., Judkins, D., Lo, A., Skellan K., "Sampling from HCFA lists", ASA 1992 Joint Statistical Meetings, Boston, MA.

The Current Population Survey: Design and Methodology. Technical Paper 40, January 1978, Department of Commerce, Bureau of the Census.

Banks, M., Shapiro, G., "Variance of the Current Population Survey, Including Within- and Between - PSU Components and the Effect of the Different Stages of Estimation," Proceeding of Social Statistics Section, ASA Annual Meeting, 1971.

Gunlicks, C.A., Corteville, J.S., Manseur, K., "Current Population Survey Variance Properties", Proceedings of the Section on Survey Research Methods, 1997.

Huang, E.T., Ernst, L.R., "Comparison of an Alternative Estimate to the Current Composite Estimate in CPS,: Proceedings of Section on Survey Research Methods, ASA Annual Meeting, 1981.

Table 1. MCBS "backward longitudinal" samples

		Relevant data									
		Access to care						Cost and use			
Year*	Subpopulation**	91	92	93	94	95	96	92	93	94	95
One-Year											
1992	Beneficiaries eligible on or before 1/1/91	X	X					X			
1993	Beneficiaries eligible on or before 9/1/92		X	X					X		
1994	Beneficiaries eligible on or before 12/31/92			X	X					X	
1995	Beneficiaries eligible prior to 1/1/94				X	X					X
1996	Beneficiaries eligible on or before 1/1/95					X	X				
Two-Year											
1993	Beneficiaries eligible on or before 1/1/91	X	X	X				X	X		
1994	Beneficiaries eligible on or before 9/1/92		X	X	X				X	X	
1995	Beneficiaries eligible on or before 12/31/93			X	X	X				X	X
1996	Beneficiaries eligible prior to 1/1/94				X	X	X				X
Three-Year											
1994	Beneficiaries eligible on or before 1/1/91	X	X	X	X			X	X	X	
1995	Beneficiaries eligible on or before 9/1/92		X	X	X	X			X	X	X
1996	Beneficiaries eligible on or before 12/31/93			X	X	X	X		X	X	X

* "Year" indicates the Access to Care File containing the weights for each "backward longitudinal" sample.

** Each full Access to Care sample represents the "always eligible" population for the relevant year.

Figure 1

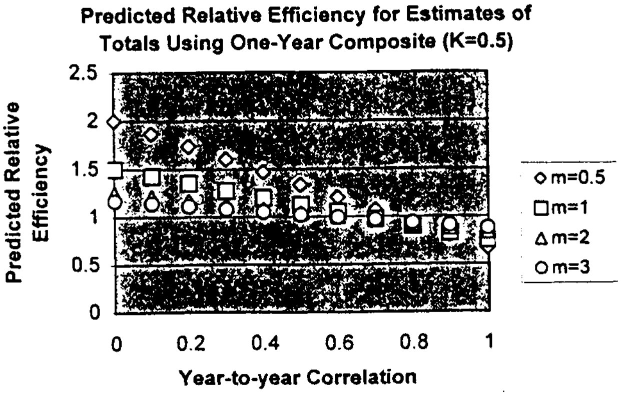
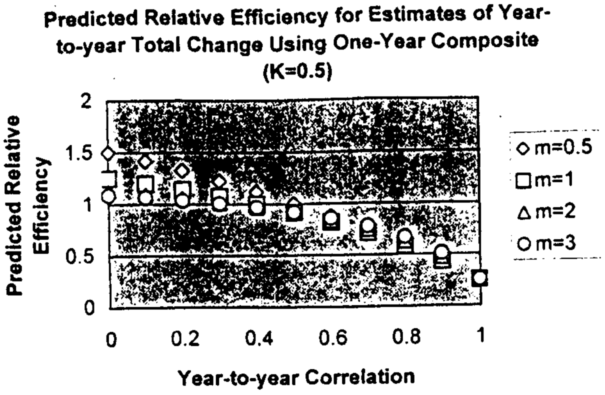


Figure 2



$$m = \frac{\text{size of overlap sample}}{\text{size of non - overlap sample}}$$

Table 2. Variance efficiency relative to simple estimate from 1995 cost and use for one-year composite estimates (a)

	Analysis domain							
	HMO		NonMetro		85+		Medicaid	
Variable/response category	1995 Estimate	1994-1995 Change (b)	1995 Estimate	1994-1995 Change (b)	1995 Estimate	1994-1995 Change (b)	1995 Estimate	1994-1995 Change (b)
Demographic characteristics:								
Income <= 25K	1.157	0.994	1.054	1.041	0.907	0.902	0.950	0.576
Education: 1-8	1.070	0.726	0.962	0.523	0.809	0.716	0.968	0.598
Marital status: M	0.981	0.669	0.863	0.590	0.869	0.677	1.077	0.636
Health status:								
General health: fair, poor	1.147	1.013	1.018	1.111	0.816	0.832	0.928	0.787
Health limited activity: most, all of the time	1.156	1.093	0.992	0.936	1.008	0.951	1.103	1.064
Difficulty walking	0.991	0.990	0.981	0.848	0.993	0.946	1.094	1.005
Difficulty shopping	1.466	1.535	1.112	0.791	1.078	0.918	1.051	0.986
Hypertension	1.061	0.730	1.004	0.540	0.873	0.762	0.977	0.701
Total CY charges:								
Inpatient	1.102	1.050	0.908	0.930	1.154	1.031	1.305	1.216
Outpatient	1.111	1.195	1.161	1.077	1.554	1.304	1.007	0.973
Physician/supplier	1.126	1.706	0.799	0.951	1.706	1.694	1.043	0.913
Home health	1.242	1.169	0.912	0.754	0.918	0.842	1.587	0.407
Prescription medicines	1.142	1.001	0.941	0.717	0.963	0.916	1.064	0.959
Total CY reimbursements:								
Inpatient	1.074	1.057	0.922	0.964	1.156	0.167	1.236	1.151
Outpatient	1.259	1.296	1.107	1.010	1.414	1.186	1.052	0.950
Physician/supplier	1.003	0.970	0.870	0.882	1.499	1.484	0.962	0.913
Home health	1.242	1.175	0.892	0.703	0.938	0.754	0.831	0.732

(a)Values in the table are ratios of the variance estimate for the composite estimate to the variance estimate for the simple estimate. Ratio less than 1 indicate improvement with compositing.

(b)Estimates of change are based on difference between the CY 1995 estimate (with or without compositing) and the (uncomposited) CY 1994 estimate.