Variance Estimation and the Components of Variance for the Medicare Current Beneficiary Survey Sample[*]

Annie Lo and Adam Chu
Westat

Keywords: Multistage probability sample, components of variance, generalized variance function, Medicare population

1. Introduction

Since its inception in 1991, the Medicare Current Beneficiary Survey (MCBS) has been conducted by the Centers for Medicare and Medicaid Services (CMS), Department of Health and Human Services, to collect timely and policy-relevant data on access to health care, health status, source of care, health care utilization and costs, satisfaction with health care, and other health-related topics (Sharma, Chan, Liu, and Ginsberg, 2001). The survey is conducted with random samples of Medicare beneficiaries residing in the United States and Puerto Rico. A new sample (referred to as a "panel") is selected for the MCBS each year using a stratified multistage probability sample design. Judkins and Lo (1993) developed the variance estimation procedures for the MCBS, computed estimates of the contributions to variance associated with the different stages of sampling, and constructed generalized variance functions for selected statistics. The results presented in that paper were based on the initial MCBS sample consisting of a single large panel. Since the publication of those results, the MCBS sample has evolved from a fixed panel to a rotating panel design and was updated (redesigned) in 2001 (Lo, Chu, and Apodaca, 2002). The purpose of this paper is to extend the earlier work by providing more up-to-date examples of variance components and generalized variance functions, and further discussion of the use of variance components for future design purposes.

2. The MCBS Sample Design

The MCBS employs a stratified multistage probability sample design with three stages of selection. The first stage involves the selection of primary sampling units (PSUs) consisting of MSAs (metropolitan statistical areas) and groups of rural (nonMSA) counties. The PSU sample was originally designed and selected in 1991 (Apodaca, Judkins, Lo, and Skellan, 1992). Up until 2000, all of the new beneficiary samples (panels) were selected from the same PSUs. However, the continued use of the original PSU sample resulted in losses in both sampling precision and operational efficiency. In 2000, based on an evaluation of the existing PSU sample design, a decision was made to reselect the PSUs. The analyses leading to that decision and the procedures used to update and select the new MCBS PSU sample are reported in Lo, et al (2002). Like the original sample, the new PSUs were selected with probabilities proportionate to population within strata defined by Census region, metropolitan status, and selected PSU-level socio-economic characteristics. Thus, beginning with the selection of the 2001 panel, all of the subsequent beneficiary samples have been selected from the redesigned PSUs.

The second sampling stage consists of the selection of ZIP Code areas within the sampled PSUs. To facilitate linking with available county-level data, the second-stage sampling unit is defined to be the part of the ZIP Code area that is physically contained within a given county. In other words, ZIP Code areas that cross county borders are subdivided by county into separate units called "ZIP fragments." For sampling purposes, small ZIP fragments are combined into clusters where necessary to ensure that each ZIP cluster would provide a reasonable workload for interviewers if selected for the sample. At the third and final stage of selection, beneficiaries within the sampled ZIP clusters are stratified by age and subsampled at rates designed to yield self-weighting (equal probability) samples of beneficiaries within each of seven age groups. In general, the relative overall sampling rates specified for the MCBS have ranged from a low of 1 for the 70-to-74 year-old age group to about 4 for the under 45 year-old age group.

The MCBS was originally intended to be a true longitudinal survey in which the sampled Medicare

[*] The authors gratefully acknowledge Frank Eppig at CMS for his support, Richard Sigman for his review of the paper, and Josephine Auguste for her programming support. The opinions expressed in this paper are those of the authors, and do not represent CMS.

beneficiaries would be interviewed three times a year throughout the remainder of their lives. However, after two years of data collection, it became clear that this would be impracticable. Thus, a decision was made to switch from a fixed panel design to a rotating panel design in which roughly one-third of the existing sample (i.e., the oldest panel) is retired each year, and a new panel is selected to replace it. Under this design, beneficiaries in each newly selected panel are interviewed three times a year (in periods roughly corresponding to winter, summer, and fall) for a maximum of four years. The general rotation scheme is diagrammed in Figure 1 where it can be seen that a new panel is introduced in the fall round of each year, remains in the study for four years, and then is released after 12 rounds of data collection. For example, the 1999 panel was introduced in fall of 1999 and was released prior to the fall of 2003. Additional details about the rotating panel design are given in Lo, et al (2002).

In 1991, over 15,000 beneficiaries were selected for the initial round of the MCBS. In each of the following two years, supplemental samples of about 2,400 beneficiaries per year were added to the original sample to compensate for sample attrition and to give coverage to newly enrolled Medicare beneficiaries. With the implementation of the rotating panel design in 1994, however, the number of beneficiaries selected for each annual supplement (i.e., nationally representative panel) has been between 6,300 and 6,500 beneficiaries per year.

1999			2000			2001			2002			2003		
W	S	F	W	S	F	W	S	F	W	S	F	W	S	F
1995	1995													
1996	1996	1996	1996	1996										
1997	1997	1997	1997	1997	1997	1997	1997							
1998	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998				
		1999	1999	1999	1999	1999	1999	1999	1999	1999	1999	1999	1999	
					2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
								2001	2001	2001	2001	2001	2001	2001
											2002	2002	2002	2002
														2003

W = winter (January to April); S = summer (May to August); F = fall (September to December).

Figure 1. Assignment of panels to MCBS data collection rounds by year

3. Access-to-care Data Files

CMS publishes MCBS data from each fall interview round through periodic releases of its Access to Care data files[†]. To date, CMS has released Access to Care data for calendar years 1991 through 2003. The Access to Care data files typically contains data for over 16,000 Medicare beneficiaries in four MCBS panels (e.g., see last column in Figure 1). The data included in the Access to Care releases are primarily based on interviews conducted during the most recent fall interview round (September through December of the specified calendar year). These items include data on access to health care, satisfaction with care and usual source of care, demographic characteristics, health insurance, and health status and functioning. In the Access to Care data files, the information collected in the survey is augmented with data on the use and program cost of Medicare services from Medicare claims data files.

In this paper, we used the 2003 Access to Care data file to develop variance estimates for selected data items in the Health Status and Functioning component of the fall round interview. The Health Status and Functioning component contains information about the sample beneficiary's health status, including self-reported height and weight, a self-assessment of vision and hearing, use of preventive measures such as immunizations and mammograms, smoking status, and history of illnesses and medical conditions. Standard measures of activities of daily living (ADLs) and instrumental activities of daily living (IADLs) are also available in the file.

The beneficiaries included in the 2003 Access to Care data file consist of a cross-section of all beneficiaries who were enrolled in Part A or Part B of the Medicare program throughout calendar year 2003. This "always enrolled" population includes beneficiaries enrolled in Medicare as of January 1, 2003 who were alive at the time of the fall interview. The 2003 Access to Care data file contains 16,003 beneficiaries in four panels (see Figure 1).

Also included in the Access to Care data files are a series of survey weights. These include weights for cross-sectional analyses of the Access to Care data file and three sets of weights for longitudinal analyses. As described below, only the cross-sectional weights are used in the analysis presented in this paper. Information about the various types and uses of the longitudinal weights are given in Ferraro and Liu (2005). Associated with each set of full-sample analytic weights are 100 replicate weights. Together, the full-sample and associated replicate weights are used to compute weighted MCBS estimates and their variances. The cross-sectional sampling weights developed for analysis of the 2003 Access to Care data have been adjusted to

[†] Researchers interested in obtaining Access to Care data are required to sign a data use agreement. For information about the availability of these files, visit CMS's website: www.cms.hhs.gov/MCBS.

compensate for nonresponse, undercoverage, and overlapping coverage of the constituent panels. Additional details about the MCBS weighting process are given in Judkins and Lo (1993) and in Lo, et al (2002).

4. Components of Variance

Approximately unbiased estimates of population totals derived from the MCBS can be expressed as weighted sums of the form:

$$\hat{y}_g = \sum_{p=1}^{P} \sum_{i=1}^{n_p} w_{pi} y_{pi} , \tag{4.1}$$

where y_{pi} is the observed value of the characteristic being estimated for the i th sampled beneficiary in panel p and w_{pi} is the corresponding sampling weight. As noted in Section 3, estimates derived from Access to Care data releases are typically based on $P = 4$ panels (where the index p is treated as an ordinal variable ranging from oldest to newest). However, the panels differ somewhat in coverage. In general, if C_p denotes the Access to Care population represented by panel p, then $C_p \supset C_{p-1}$ for $p = 2,3,4$. The subsets, $G_1 = C_1$ and $G_g = C_g \cap C_{g-1}^c$ $(g = 2,3,4)$ are referred to as "combination groups." The combination groups define a partition of C_4 (the population covered by the Access to Care file) where G_g represents that part of C_4 that is represented by exactly 5-g panels in the MCBS sample. Specifically, G_1 is the population of surviving Medicare beneficiaries who enrolled as of year 1 (corresponding to panel 1), G_2 is the population of surviving Medicare beneficiaries who enrolled between years 1 and 2, G_3 is the population of surviving Medicare beneficiaries who enrolled between years 2 and 3, and G_4 is the population of surviving Medicare beneficiaries who enrolled between years 3 and 4. Note that the expected values of estimates derived from the different panels are equal for a particular combination group.

To bring out these aspects of the MCBS estimator, $\hat{y}$ can be rewritten as:

$$\hat{y} = \sum_{g=1}^{G} \hat{y}_g = \sum_{g=1}^{G} \sum_{p=1}^{P} a_{gp} \hat{y}_{gp} , \tag{4.2}$$

where

$$\hat{y}_{gp} = \sum_{i=1}^{n_{gp}} w_{gpi}^{NR} y_{gpi} \tag{4.3}$$

is the estimated total for the g th combination group. The sampling weights, w_{gpi}^{NR}, in equation (4.3) are panel-specific nonresponse-adjusted weights that inflate the survey results for panel p to population levels; thus,

$$\hat{y}_g = \sum_{p=1}^{P} a_{gp} \hat{y}_{gp} \tag{4.4}$$

is a composite estimate of the population total for the g th combination group based on P panels. For combination group g, the a_{gp} 's in equation (4.2) are roughly proportional to the corresponding panel sample sizes, n_{gp} and are subject to the condition that $a_{g1} + a_{g2} + \ldots + a_{gp} = 1$.[‡]

The reason for decomposing the MCBS estimator is the following. The variance components analysis originally presented in Judkins and Lo (1993) was based on data from a single large sample (i.e., the initial MCBS panel). Under that design, the variance of an estimated total can be expressed approximately[§] as the sum of three components:

$$\mathrm{var}(\hat{y}) = \frac{\sigma_1^2}{m_{NSR}} + \frac{\sigma_2^2}{m\bar{n}} + \frac{\sigma_3^2}{m\bar{n}\bar{\bar{q}}} , \tag{4.5}$$

where σ_1^2 is the between-PSU unit variance of the characteristic of interest among the NSR PSUs; σ_2^2 is the between-ZIP-cluster unit variance; σ_3^2 is the average within-ZIP-cluster unit variance; m_{NSR} is the number of non-self-representing (NSR) PSUs in the sample; m is the total number of PSUs in the sample; $\bar{n}$ is the average number of sample ZIP clusters per PSU; and $\bar{\bar{q}}$ is the average number of beneficiaries per ZIP cluster.

[‡] The actual combination groups used to derive the MCBS weights were defined by both year of enrollment and age group (see Lo, Chu, and Apodaca, 2002). However, the combination factors typically do not vary greatly by age group, and so this added complexity is ignored in the above discussion.

[§] This formula is oversimplified, but serves to illustrate the main points. It is based on equation (14.1) on page 397 of the book by Hansen, Hurwitz, and Madow (1953).

The three terms in equation (4.5) represent the contributions to variance due to sampling PSUs, ZIP clusters within PSUs, and beneficiaries within ZIP clusters, respectively. Note that the self-representing (SR) PSUs do not contribute to the between-PSU component of variance. Analysis of the variance components given in equation (4.5) can be useful for future design purposes. Because total sampling variances can be decomposed into components corresponding to sampling stages, the estimates of the components of variance can suggest more optimal choices for the sample design at each stage.

However, under the rotating panel design currently employed in the MCBS, the decomposition of variances becomes more complicated. For example, consider an estimated total for a particular combination group g . Then from equation (4.4),

$$\mathrm{var}\left(\hat{y}_g\right) = \mathrm{var}\left(\sum_{p=1}^{4} a_{gp}\hat{y}_{gp}\right) = \sum_{p=1}^{4} a_{gp}^2\, \mathrm{var}\left(\hat{y}_{gp}\right) + 2\sum_{p>\ q}\sum a_{gp}a_{gq}\, \mathrm{cov}\left(\hat{y}_{gp}, \hat{y}_{gq}\right). \qquad (4.6)$$

Replacing $\mathrm{var}\left(\hat{y}_{gp}\right)$ and $\mathrm{cov}\left(\hat{y}_{gp}, \hat{y}_{gq}\right)$ by formulas such as (4.5) then yields

$$\mathrm{var}\left(\hat{y}_g\right) = \sum_{p=1}^{4} a_{gp}^2 \left[\frac{\sigma_{1p}^2}{m_{NSR}} + \frac{\sigma_{2p}^2}{m\bar{n}} + \frac{\sigma_{3p}^2}{m\bar{n}\bar{\bar{q}}_p}\right] +$$

$$2\sum_{p>q}\sum a_{gp}a_{gq}\left[\frac{\sigma_{1pq}}{m_{NSR}} + \frac{\sigma_{2pq}}{m\bar{n}} + \frac{\sigma_{3pq}}{m\bar{n}\bar{\bar{q}}_q}\right], \qquad (4.7)$$

where σ_{1p}^2 is the between-PSU unit variance of the characteristic of interest among the NSR PSUs for panel p ; σ_{2p}^2 is the average (within PSU) between-ZIP-cluster unit variance for panel p ; σ_{3p}^2 is the average within-ZIP-cluster unit variance for panel p ; σ_{1pq} is the between-PSU unit covariance between panel p and panel q among the NSR PSUs; σ_{2pq} is the between-ZIP cluster unit covariance between panel p and panel q ; and σ_{3pq} is the within-ZIP cluster unit covariance between panel p and panel q ; m_{NSR} , m and $\bar{n}$ are defined as in equation (4.5) but $\bar{\bar{q}}_p$ and $\bar{\bar{q}}_q$ are now the average cluster sizes for panel p and q , respectively.

Note that m_{NSR} and $\bar{n}$ can be assumed to be constant if all of the panels included in the Access to Care release are selected within the same PSUs and (essentially) the same ZIP clusters within PSUs. The three unit variances σ_{1p}^2 , σ_{2p}^2 and σ_{3p}^2 can similarly be treated as constant and independent of p . Since the between-PSU and between-ZIP cluster unit covariances measure year-to-year covariability within the same PSUs and ZIP clusters, it would not be unreasonable to assume that $\sigma_{1pq} = \sigma_1^2$ and $\sigma_{2pq} = \sigma_2^2$. Finally, we note that within the sampled ZIP clusters, the beneficiary samples selected for different panels may be considered independent. In this case, equation (4.7) can be further simplified by assuming $\sigma_{3pq} = 0$.

Under these simplifying conditions, the variance reduces to

$$\mathrm{var}\left(\hat{y}_g\right) = \left(\sum_{p=1}^{4} a_{gp}\right)^2 \left[\frac{\sigma_1^2}{m_{NSR}} + \frac{\sigma_2^2}{m\bar{n}}\right] + \sum_{p=1}^{4} a_{gp}^2 \frac{\sigma_3^2}{m\bar{n}\bar{\bar{q}}_p} \qquad (4.8)$$

Finally, we note that for each g , $\sum_{p=1}^{4} a_{gp} = 1$. Moreover, since the panel sample sizes are roughly equal, the last term in (4.8) is approximately $\sigma_3^2 \left(m\bar{n}\bar{\bar{q}}\right)^{-1}$. Thus, while the complex form of the variance components given by (4.7) provides a more general description of the structure of MCBS variances, the simpler form (4.5) provides a useful and appropriate approximation for design purposes. Thus, in what follows, the simpler model given by (4.5) will be used to examine alternative designs.

In our analysis of variance components, it is useful to express variances in relative terms. The relative variance (relvariance) of a sample-based estimate is defined to be the variance of the estimate divided by the square of the quantity being estimated. The square root of the relvariance is the coefficient of variation (i.e., relative standard error) of the estimate. Corresponding to equation (4.5), the relvariance of an estimated total, $\hat{y}$, can be expressed as:

$$V_{\hat{y}}^2 = \frac{R_1^2 V_{1NSR}^2}{m_{NSR}} + \frac{V_2^2}{m\bar{\bar{n}}} + \frac{V_3^2}{m\bar{\bar{n}}\bar{\bar{q}}} \tag{4.9}$$

where $R_1 = Y_{NSR} / (Y_{NSR} + Y_{SR})$, Y_{NSR} = the total value of the quantity being estimated that is included in the NSR PSUs, Y_{SR} = the total value of the quantity being estimated that is included in SR PSUs, V_{1NSR}^2 is the between-PSU unit relvariance among the NSR PSUs; V_2^2 is the between-ZIP-cluster unit relvariance; V_3^2 is the average within-ZIP-cluster unit relvariance.

5. Estimation of Total Variances

A form of balanced repeated replication (BRR) referred to as Fay's method has been used to compute the variance of estimates derived from the MCBS. The idea behind replication is to select subsamples (replicates) from the full sample, calculate the statistic of interest for each replicate, and then use these replicate statistics to estimate the variance of the full-sample statistic. The variance estimates calculated using Fay's method account for complex features of the sample design such as clustering by PSU and ZIP Code, stratification, unequal probabilities of selection, and nonresponse weighting adjustments. Fay's estimate of the relvariance is given by

$$V^2\left(\hat{\theta}\right) = \frac{1}{(1-k)^2} \left(\frac{1}{G} \sum_{g=1}^{G} \left(\hat{\theta}_{(g)} - \hat{\theta} \right)^2 \right) \frac{1}{\hat{\theta}^2} \tag{5.1}$$

where $\hat{\theta}$ is the parameter estimate based on the full sample, $\hat{\theta}_{(g)}$ is the g^{th} replicate estimate of θ based on the observations included in the g^{th} replicate, G is the total number of replicates formed, and 100(1-*k*) percent is a constant referred to as Fay's perturbation factor. For the MCBS, a value of 0.3 was chosen for *k* corresponding to a perturbation factor of 70 percent (Judkins, 1990).

For estimation of total variances that account for all three sampling stages, the required replicates were created as follows. First, each of the 39 strata defined for sampling NSR PSUs was treated as a separate variance-estimation stratum. Each of the two sampled PSUs within these strata was specified as variance-estimation units. An additional 61 variance strata were created for the SR PSUs by (a) sorting ZIP clusters in sample selection order within each PSU, (b) forming pairs or triplets of consecutive ZIP clusters within each PSU, (c) labeling the ZIP clusters within each pair or triplet as 1, 2, or 3, and (d) systematically assigning the pairs or triplets to one of 61 variance strata. Within each of the variance strata constructed in this manner, all of the ZIP clusters labeled "1" constituted variance unit 1, those labeled "2" constituted variance unit 2, and so on. Thus, for estimating the total variance given by (5.1), variance strata were created at the PSU level for the NSR PSUs and at the ZIP cluster level for the SR PSUs. One hundred replicates were then created by selecting one of the two variance units from each variance stratum based on a Hadamard matrix of 1s and -1s (e.g., see McCarthy, 1966). Thus, each of the 100 replicates is a balanced half sample that mirrors the design of the full MCBS sample.

Use of formula (5.1) requires separate weights for each of the replicates. To form the weights for the replicate estimates, the full-sample weights of the units included in the replicate are multiplied by a factor k $(0 \le k \le 1)$ while the full-sample weights of the remaining half are multiplied by 2-*k*.

A second version of replicate weights was also developed to estimate the total within-PSU variance. This was accomplished by defining the variance strata and variance units somewhat differently. For the SR PSUs, the same variance strata and variance units defined for total variances were used. For the NSR PSUs, the variance strata were created at the ZIP cluster level. ZIP clusters were sorted in the order in which they had been selected, and variance strata were formed by pairing consecutive ZIP clusters. With this alternative definition of variance strata, another version of replicate weights designed for the estimation of total within-PSU variances was created.

Finally, a third version of replicate weights designed to reflect just the within-ZIP cluster sampling was created. This was accomplished by sorting the sample of beneficiaries in the order in which they were selected, and then forming variance strata and variance units by systematically pairing beneficiaries in the sorted list. With this version of variance strata, a third version of replicate weights for the estimation of total within-ZIP cluster variances was created.

Using the three alternative sets of replicate weights defined above, three versions of the variance of a sample-based statistic were computed. Let v_{T1}^2 denote the relvariance computed from formula (5.1) using the first version of replicate weights. Let v_{T2}^2 denote the relvariance computed from formula (5.1) using the second version of replicate weights. Finally, let v_{T3}^2 denote the relvariance computed from formula (5.1) using the third version of replicate weights. The total between-PSU relvariance, v_{BP}^2 was then estimated as $v_{BP}^2 = v_{T1}^2 - v_{T2}^2$. Note that v_{BP}^2 corresponds to the first of the three terms in equation (4.9).

Similarly, the between-ZIP cluster relvariance, v_{BZ}^2 was estimated as $v_{BZ}^2 = v_{T2}^2 - v_{T3}^2$. Note that v_{BZ}^2 corresponds to the second of the three terms in equation (4.9). Finally, v_{T3}^2, the total within-ZIP cluster relvariance, corresponds to the last term in equation (4.9).

We computed the relative variance components, v_{BP}^2, v_{BZ}^2, and v_{T3}^2 for 377 prevalence estimates in the Health Status and Functioning file of the 2003 Access to Care data release. These estimates are based on the three most recent panels in the 2003 Access to Care data file. The 2000 panel was excluded from the estimates because it is based on the older MCBS design. Table 1 summarizes the results for selected groups of statistics. These groups of statistics correspond to related groups of questions in the Health Status and Functioning portion of the interview and include: activities of daily living (ADLs related to bathing or showering, dressing, etc.); falls, (whether the SP has fallen down in the past year, needed medical attention for a fall, etc.) health (general health); instrumental activities of daily living (IADLs related to using the telephone, doing light and heavy housework, etc.); medical conditions (heart disease, high blood pressure, etc.); memory loss; preventive health care measures (flu shots, blood pressure checked, etc.); smoking status, and questions related to hearing, vision, and teeth.

Table 1. Relative variance estimates for groups of statistics in the 2003 MCBS Access to Care file

Group of health status and functioning statistics	Total relative variance	Percent contribution		
		Between PSU	Between ZIP cluster	Within ZIP cluster
ADL (low prevalence: ≤ 1%)	0.04160	9.7%	3.1%	87.2%
ADL (medium prevalence: > 1% - ≤ 5%)	0.00479	13.4	4.8	81.8
ADL (high prevalence: > 5%)	0.00100	9.4	3.0	87.6
Difficulty doing activities	0.00114	9.7	4.1	86.2
Falls	0.00111	12.9	5.2	81.9
Health	0.00102	11.3	4.0	84.7
Hearing	0.01339	15.7	10.9	73.4
IADL	0.00309	26.9	5.8	67.3
Incontinence	0.00327	11.4	3.9	84.7
Medical conditions	0.00578	10.3	6.8	82.9
Memory loss	0.00089	13.4	4.9	81.7
Preventive health care measures	0.01095	25.2	1.8	73.0
Smoking	0.00715	21.4	0.2	78.4
Teeth	0.00048	0.0	27.4	72.6
Vision	0.00343	13.9	1.3	84.7

The total relvariances shown in Table 1 are the average values of the relvariances for a particular group of health status and functioning statistics, and the corresponding contributions to relvariance are expressed as percentages of the total relvariance. Due to the large number of ADL statistics and variation in prevalences, we further subdivided the ADL statistics into low, medium, and high prevalence categories. The major source of variance for statistics related to health status and functioning is largely due to the within-ZIP cluster sampling. However, for some statistics, the variance due to sampling PSUs accounts for 25 percent or more of the total variance. The sum of the two components of the total within-PSU relvariance ranges from about 73 percent for statistics related to IADLs to over 80 percent for statistics related to ADLS. In general, the major component of the within-PSU variance is the variance due to sampling beneficiaries within ZIP clusters. The contribution to variance due to sampling ZIP clusters within PSUs is generally 10 percent or less under the current MCBS design. It should be noted that the within-ZIP cluster variance contributions shown in Table 1 reflect the use of varying subsampling fractions by age group. For individual age groups, the within-ZIP cluster variances will be smaller.

6. Estimation of Unit Variances

The relvariance components needed for design purposes are the three unit (or "element") variances V_{1NSR}^2, V_2^2, and V_3^2 defined in equation (4.9). These unit relvariance components were estimated as follows:

$$V_{1NSR}^2 = \frac{v_{BP}^2 m_{NSR}}{R_1^2} \tag{6.1}$$

$$V_2^2 = v_{BZ}^2 m\bar{n} \tag{6.2}$$

$$V_3^2 = v_{T3}^2 m\bar{n}\bar{\bar{q}} \tag{6.3}$$

where v_{BP}^2, v_{BZ}^2, and v_{T3}^2 are the relvariance contributions described in Section 5.

Table 2. Estimated unit relvariances for groups of statistics in the 2003 MCBS Access to Care file

Group of health status and functioning statistics	Between PSU	Between ZIP cluster	Within ZIP cluster
ADL (low prevalence: ≤ 1%)	0.656	1.636	465.97
ADL (medium prevalence:>1% - ≤ 5%)	0.104	0.295	50.25
ADL (high prevalence: > 5%)	0.015	0.039	11.30
Difficulty doing activities	0.018	0.060	12.61
Falls	0.023	0.074	11.69
Health	0.019	0.053	11.03
Hearing	0.342	1.868	126.29
IADL	0.135	0.231	26.67
Incontinence	0.060	0.165	35.57
Medical conditions	0.097	0.507	61.58
Memory loss	0.019	0.056	9.33
Preventive health care measures	0.448	0.255	102.63
Smoking	0.249	0.022	72.00
Teeth	0.000	0.170	4.49
Vision	0.078	0.059	37.27

In equation (6.1), an empirically-based average value of 0.69 for R_1 was used. The total number of sampled PSUs was set to 107, and the number of NSR PSUs, m_{NSR} was set to 78. The average number of sample ZIP clusters per PSU and the average number of sample persons per ZIP cluster were set to $\bar{n} = 12$ and $\bar{\bar{q}} = 10$ respectively. The value of $\bar{\bar{q}} = 10$ corresponds to the sample size for the three panels that were included in the variance calculations. Table 2 summarizes the estimated unit relvariances for selected groups of statistics. Since the components are expressed in terms of relvariances, the magnitudes of the unit relvariances depend on the prevalence of the item being estimated. As can be seen in the table, low prevalence items such as low-prevalence ADLs, hearing-related items, and items on preventative measures have large relvariance components. In all cases, the within-ZIP cluster unit relvariance is the dominant component.

7. Alternative Designs

To investigate alternative sample designs, both variance and costs need to be considered. Formal cost-variance analyses require the estimation of a cost function that breaks out the total cost into separable components associated with the different stages of sampling. An example of simple cost function is:

$$C = C_1 m + C_2 m\bar{n} + C_3 m\bar{n}\bar{\bar{q}} \, , \tag{7.1}$$

where C is the total cost exclusive of fixed overhead costs, C_1 is the unit cost per PSU in the sample, C_2 represents the unit cost per ZIP cluster; and C_3 represents the unit cost per beneficiary. Although the above model greatly oversimplifies the complex MCBS cost structure, it serves to illustrate the possible cost implications associated with alternative designs.

For the alternative designs considered below, we assume that m is fixed at 107. Thus, the first term in equation (7.1) can be absorbed in the fixed overhead costs. A cost function that reflects only the costs associated with sampling ZIP clusters and beneficiaries, C_w, is then given by:

$$C_w = C_2\bar{n} + C_3\bar{n}\bar{\bar{q}} \, . \tag{7.2}$$

In the MCBS, the unit cost associated with a sampled ZIP cluster is much larger than the unit cost associated with a sampled beneficiary. For example, it has been estimated that increasing the average number of sampled ZIP clusters per PSU from the current 12-13 ZIP clusters to 17 (and assuming the same $\bar{\bar{q}}$), will increase total costs by roughly 20 percent. The reason for the increase is that the unit cost per sampled ZIP cluster, which is dominated by interviewer travel costs, is significantly higher than the unit cost per beneficiary. An approximate cost function that reflects these assumptions is given by equation (7.2) where $C_2 = 0.04C_w$ and $C_3 = 0.0033C_w$. Under this cost model, we obtain the following relationship between $\bar{n}$ and $\bar{\bar{q}}$ for designs of approximately equal cost: $\bar{\bar{q}} = \left(\frac{1}{\bar{n}} - 0.04\right)\left(\frac{1}{0.0033}\right)$.

Using the unit relvariances in Table 2 and equation (4.9), we estimated the relvariances for particular values of $\bar{n}$ and $\bar{\bar{q}}$ corresponding to the three design alternatives specified in Table 3.

Table 3. Total relvariances for three designs

Group of health status and functioning statistics	Design 1	Design 2	Design 3
	$\bar{n}$ =10, $\bar{\bar{q}}$ = 18	$\bar{n}$ =12, $\bar{\bar{q}}$ = 13	$\bar{n}$ =14, $\bar{\bar{q}}$ = 9
ADL (low prevalence: ≤ 1%)	0.0393	0.0441	0.0508
ADL (medium prevalence: > 1% - ≤ 5%)	0.0046	0.0051	0.0058
ADL (high prevalence: > 5%)	0.0009	0.0011	0.0012
Difficulty doing activities	0.0011	0.0012	0.0014
Falls	0.0011	0.0012	0.0013
Health	0.0010	0.0011	0.0012
Hearing	0.0135	0.0145	0.0161
IADL	0.0030	0.0033	0.0036
Incontinence	0.0031	0.0035	0.0040
Medical conditions	0.0056	0.0062	0.0070
Memory loss	0.0009	0.0009	0.0011
Preventive health care measures	0.0104	0.0115	0.0130
Smoking	0.0067	0.0075	0.0085
Teeth	0.0005	0.0006	0.0006
Vision	0.0032	0.0036	0.0042

Design 2 is essentially the current MCBS sample design. As can be seen in Table 3, the total relative variances are expected to decrease by about 10 percent if the average number of sampled ZIP clusters per PSU is decreased from 12 to 10 and the average number of sampled beneficiaries per ZIP cluster is increased from 13 to 18. This improvement in precision is achieved because (a) the between-ZIP cluster contribution does not increase appreciably with the smaller ZIP cluster sample size, and (b) the increased beneficiary sample size that can be afforded under this design more than offsets the increased between-ZIP cluster variance. On the other hand, increasing the number of ZIP clusters per PSU while maintaining the same overall survey costs will inflate variances by up to 10-15 percent for the statistics examined.

8. Generalized Variance Functions

Direct computation of the standard errors of estimates derived from the MCBS is relatively straightforward and should be standard practice. However, some analysts may not have the resources or ability to do this easily. Furthermore, direct variance estimates are themselves subject to sampling error and thus can be highly unstable (Judkins and Lo, 1993). Indirect methods of variance estimation using generalized variance functions (GVFs) provide users with an alternative and relatively simple way of estimating sampling errors. Since GVFs are based on the observed (calculated) variances for a large group of similar statistics, the resulting variance estimates tend to be more stable than direct estimates (Valliant, 1987).

In this paper, generalized variance functions (GVFs) are developed for estimates derived from the Health and Functioning section of the MCBS interview using the same methodology previously described in Judkins and Lo (1993). As mentioned earlier, the MCBS sample design has undergone some important revisions since the publication of the earlier results. Thus, the GVFs presented in this paper are expected to be more appropriate for analysis of recent Access to Care data since they are derived from samples reflecting the design modifications. The GVF is a mathematical model which describes the relationship between the relative variance of a survey estimator and its expectation. Among a number of different types of GVF models, the following model often provides a useful description of this relationship and has been used extensively (e.g., see U.S. Bureau of the Census, 1978):

$$V^2 = \frac{\sigma^2_{\hat{X}}}{X^2} = \alpha + \frac{\beta}{X}, \qquad (8.1)$$

where $\beta > 0$, V^2 is the relative variance of an estimated total $\hat{X}, \sigma^2_{\hat{X}}$ denotes the variance of $\hat{X}$, and $X = \mathrm{E}(\hat{X})$. The model described in equation (8.1) is estimated using the direct replicate-based variances discussed in Section 5. Valliant (1987) provided a theoretical justification for the model and showed that GVF relvariance estimators can perform as well or better than direct estimators in terms of bias, precision, and confidence interval construction.

The development of the GVFs involved several steps. First, the standard errors for a large number of estimates were computed using Fay's variant of BRR. Next, models were fitted to the estimates and standard errors and the parameters of these models were estimated using iterative weighted least squares. The estimated models can then be used to approximate the standard error of a similar estimate derived from the survey.

For the 2003 Access to Care data file, GVF models were estimated separately by age category, gender, and race. Note that unlike the analysis in Section 5 which used the three panels selected under the new PSU design, the estimates used to derive the GVFs were based on the entire 2003 Access to Care data file. Table 4 gives the estimates of the GVF parameters. Note that the results only apply to health and functioning statistics in the Access to Care file.

Table 4. Estimated parameters for computing generalized variances for estimates from the 2003 MCBS Access to Care Public Use File (Health Status and Functioning data)

Domain	Parameter estimates	
	a	*b*
All beneficiaries	-0.000081	3264.11
Age		
Under 45 years	-0.000680	1529.30
45 - 64	-0.000395	4046.91
65 – 69	-0.000220	3619.74
70 - 74	-0.000226	3495.00
75 – 79	-0.000119	2678.73
80 – 84	-0.000196	2191.54
85 years and older	-0.000175	1961.67
Sex		
Male	-0.000119	3021.82
Female	-0.000113	3038.11
Race		
Asian	0.000359	3141.49
Black	-0.000400	3210.26
White	-0.000082	3272.74
Other	0.001930	3299.16

Using the estimated parameters a and b given in Table 4, the standard error of an estimated total can be computed as: $se(x) = \sqrt{ax^2 + bx}$. The standard error of an estimated percentage can be computed as: $se(p) = \sqrt{\frac{bp(100-p)}{y}}$, where $p = 100\left(\frac{x}{y}\right)$, y is the total number of individuals in a particular subgroup of the population and x is the number of those individuals possessing a specified characteristic.

9. Summary

Understanding the components of variance that arise from various stages of sampling is important in designing future surveys. Together with a cost model, analysis of variance components can shed light on the relative efficiencies of alternative sample designs. Our paper illustrates a method for estimating the components of variance associated with the MCBS three-stage sample design. In practice, there is no single design that is optimum for all types of statistics. A comprehensive analysis of the costs and variances associated with alternative designs would therefore have to consider a variety of different types of statistics. In this case, the goal would be to determine a design that is roughly optimum for a broad range of statistics collected in the MCBS.

Generalized variance functions provide a simple way of estimating standard errors. The GVF models presented in this paper are appropriate for computing standard errors for means or totals of binary characteristics derived from the Health and Functioning section of the MCBS interview.

References

Apodaca, R., Judkins, D., Lo, A., Skellan, K. (1992). Sampling from HCFA lists. *Proceedings of the Survey Methods Research Section, American Statistical Association*, 250-255.

Ferraro, D. and Liu, H. (2005). Uses of the Medicare Current Beneficiary Survey for Analysis across Time. To be published in *Proceedings of the Survey Methods Research Section, American Statistical Association*.

Hansen, M., Hurwitz, W., Madow, W. (1953). *Sample Survey Methods and Theory*, New York: John Wiley & Sons.

Judkins, D. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 3, 223-240.

Judkins, D. and Lo, A. (1993). Components of variance and nonresponse adjustment procedures for the Medicare Current Beneficiary Survey, *Proceedings of the Survey Methods Research Section, American Statistical Association*, 820-825.

Lo, A., Chu, A., Apodaca, R. (2002). Redesign of the Medicare Current Beneficiary Survey Sample. *Proceedings of the Survey Methods Research Section, American Statistical Association*, 2139-2144.

Sharma, R., Chan, S., Liu, H., Ginsberg, C. (2001). Health and Health Care of the Medicare Population. *Data from the 1997 Medicare Current Beneficiary Survey.* Rockville, MD. Westat.

McCarthy, P.J. (1966). Replication: An Approach to the Analysis of Data from Complex Surveys. *Vital and Health Statistics, Series 2, No. 14.* Washington, DC: National Center for Health Statistics.

U.S. Bureau of the Census (1978). *The Current Population Survey: Design and Methodology.* Technical Paper 40, NO. C3.212:40.

Valliant, R. (1987). Generalized Variance Functions in Stratified Two-stage Sampling. *Journal of the American Statistical Association*, 82, 499-508.

Westat (2000). Medicare Current Beneficiary Survey: Evaluation of Alternative Measures of Size for Primary Sampling Unit Selection. *Technical*

Report prepared for Health Care Financing Administration, Rockville, MD.

Westat (2001). Medicare Current Beneficiary Survey: Design and Selection of the 2001 PSU Sample. *Technical Report prepared for Health Care Financing Administration*, Rockville, MD.

Wolter, K. (1985). *Introduction to Variance Estimation*, New York: Springer-Verlag.