SAMPLING FROM HCFA LISTS

Rick Apodaca, Westat, Inc., David Judkins, Westat, Inc., Annie Lo, Westat, Inc.,
Kim Skellan, HCFA
Westat, Inc. 1650 Research Blvd. Rockville, MD

Key Words: Survey design, Frame, Response rate

1. Introduction

A large sample of Medicare beneficiaries was recently drawn for the Medicare Current Beneficiary Survey (MCBS) from lists maintained by the Health Care Financing Administration (HCFA). Two HCFA lists were used in sampling: the 1988 Continuous Medicare History Sample File (CMHS) was used in selecting ZIP code clusters in the sample primary sampling units (PSUs); and the Health Insurance Master File (HIM) was used in selecting Medicare beneficiaries in the sample ZIP code clusters.

The MCBS is a continuous, multi-purpose panel survey of Medicare beneficiaries. Included in the study are the aged and the disabled residing in households and nursing homes. Considerable study was conducted to assess the possibility of oversampling nursing home residents. Due to the lack of good sensitivity of HCFA and SSA indicators of institutionalization, the plan to oversample nursing home residents was abandoned. A decision was reached to oversample the oldest old which would result in a modest oversample of nursing home residents.

This paper presents the sample design of the MCBS, advantages of clustering, and some of the difficulties encountered in sampling. The paper also examines coverage issues, address quality, and response rates.

2. Overview of Sampling

The MCBS sample design is a stratified area probability design with three stages of selection: (1) selection of 107 primary sampling units (PSUs); (2) selection of 1,163 ZIP clusters within the sample PSUs; (3) selection of 15,215 Medicare beneficiaries within the sample ZIP clusters and PSUs. The sample size was designed to yield complete annual data on 12,000 beneficiaries.

The first stage of selection was the sampling of PSUs. The 1981 Westat general purpose sample of 100 PSUs was expanded to 107 PSUs for the MCBS. This general purpose sample was expanded rather than replaced because of the availability of experienced interviewers. The additional PSUs brought some southern and western metropolitan statistical area (MSAs) into the sample that had experienced significant increases in the elderly population during the 1980s. The PSUs are composed of MSAs and clusters of non-metropolitan counties. Within region and metropolitan status, PSUs were grouped into strata defined to be internally homogeneous with respect to socio-economic data and to be roughly equal in size. The strata covered the 50 states, the District of Columbia, and Puerto Rico. Measures of size were mostly based on 1980 population. The measure of size was updated for those areas that had absorbed significant numbers of the elderly during the 80s. Large metropolitan areas such as New York and Los Angeles constituted their own strata and were selected with certainty. From each of the non-certainty strata, two PSUs were selected with probability proportionate to 1980 population.

The second stage of sampling was the random selection of ZIP clusters from within the populous sample PSUs. ZIP codes that cross county borders were split by county. The resulting pieces were called ZIP fragments. A measure of size was assigned to each ZIP fragment from a summary of the 5-percent 1988 Continuous Medicare History Sample File (CMHS). The measure of size was closely related to the total count of Medicare beneficiaries residing in the ZIP fragment, but beneficiaries in domains to be oversampled (such as disabled persons under age 65) were counted more heavily than persons to be undersampled (such as persons aged 66 to 69). Some of the ZIP fragments had very small number of beneficiaries residing in them. These small ZIP fragments were collapsed with each other or with large ZIP fragments until a reasonable aggregate measure of size had been achieved for each cluster, yeilding 15,102 ZIP clusters. A sample of 1,163 ZIP clusters was selected with probability proportionate to the measure of size using systematic sampling with a random start.

The selection of beneficiaries constituted the third stage of sampling. Two steps were involved in this sampling process. In late 1990, a preliminary systematic sample of 27,773 beneficiaries was selected from the 5-percent sample of the Health Insurance Master File (HIM) with probability proportionate to size. A measure of size was defined to make the beneficiary sample as close as possible to self-weighting within each of the age domains (aged 0-44, 45-64, 65-69, 70-74, 75-79, 80-84, 85+).

The preliminary sample only covered Medicare beneficiaries of record as of September 1990. New beneficiaries who enrolled after August 1990 but on or before January 1, 1991 was added to the preliminary sample. In June 1991, the preliminary sample was trimmed to 15,215 beneficiaries for the first round of interviewing[1] .

3. Advantages and Difficulties of Clustering

Clustering introduces important operational efficiencies into surveys that involve face-to-face interviewing. Simultaneously, it decreases precision as compared to an unclustered sample of the same sample size. For MCBS, it also led to difficulties in coverage. This paper focuses on design effects and on the problem of maintaining coverage. The operational efficiencies were not quantified, but they are thought to have been important in keeping the cost per case below early projections.

MCBS uses resident interviewers for the most part. It was desired that the number of PSUs be small enough so that there would be sufficient work for one or two interviewers in each PSU but large enough to keep between-PSU variance small.

In an area sample, it is almost always clear whether a person lives in a sample PSU. The situation was much less clear when sampling from HCFA lists. County codes were frequently missing or set equal to impossible values. Of the 53,899 ZIP fragments on the 5-percent 1988 CMHS file, 3.1% had county codes that were missing or invalid. Furthermore, SSA uses its own county coding system for which a firm cross-walk to FIPS codes does not appear to exist. The most aggravating feature of SSA county coding concerned blanks and zeros. There are several counties in the U.S. for which the SSA county code is either zero or blank. The representation depends on the file being used and is not always consistent within the same file.

A solution to the problem of missing/invalid county codes was to impute county based on ZIP code. This was feasible since ZIP code is seldom missing from HCFA addresses. The modal county code for each ZIP code was imputed to fragments with missing/invalid county codes with the same ZIP code. Imputation was performed successfully on 97.5% of the missing/invalid county codes.

Clustering by PSU saves a considerable amount of money, but simple random samples within PSUs can still result in a lot of expensive local travel. To further reduce costs, the sample was clustered by ZIP code within sample PSUs. The more common technique to cluster by block in area samples was considered, but block is not a variable on HCFA lists. While it is conceivable that the HCFA lists could be geocoded into decennial census blocks by the U. S. Census Bureau or by private companies such as Donnelly and R.L. Polk, the operation would have been extremely expensive, particularly since SSA/HCFA addresses do not have isolated address subfields for house number, street and place, making the geocoding very difficult.

Many county or ZIP code errors on the file were discovered. The evidence of this is the very large number of ZIP fragments (intersections of counties and ZIP codes) with very few beneficiaries. In order to be useful for clustering, units with more than a minimum number of resident beneficiaries were needed. To obtain these larger units, ZIP fragments within the same county were collapsed. The ZIP fragments in each county were first sorted by ZIP code. Starting at the top, each deficient ZIP fragment was then collapsed with succeeding fragments, backing up at the end of the county if necessary, until a minimum measure of size had been achieved for each ZIP cluster.

This clustering design was reasonable given that no prior Medicare survey experience was available from recent surveys. The MCBS was the first in-person survey of the Medicare population since the Current Medicare Survey (CMS) conducted in 1977. Future samplers may wish to consider a modification of the technique we used for clustering. That technique is to sort all beneficiaries in sample PSUs by county, mean payments for ZIP, and past individual payments, and then draw a systematic sample where the first m beneficiaries after each hit beneficiary are drawn into the sample. We did not pursue this option because of the much greater computer cost (60-70% of all beneficiaries live in sample PSUs). By restricting the person-level sampling to sample ZIP clusters, the sizes of the files to be sorted were sharply reduced. Also, this alternative technique leads to slightly looser clustering than the technique we used. (If the initial hit is near the end of a large fragment that precedes another large fragment, the alternative technique needlessly splits the cluster of persons across the two fragments.)

4. Population Covered in HCFA Lists

The Medicare program finances health care for the aged and disabled beneficiaries of the social security and railroad retirement programs and to persons requiring dialysis or a kidney transplant for end stage renal disease (ESRD). Medicare consists of two separate but complementary insurance programs: hospital insurance (HI) plan (Part A) and supplementary medical insurance (SMI) plan (Part B). HI covers inpatient hospital, some skilled nursing facilities, home health agency services and hospice care. SMI covers physicians' and related services for eligible persons who voluntarily pay premiums or whose premiums are paid for them. SMI also covers outpatient hospital services, rural health clinic visits, and home health visits.

All persons 65 years of age or over who are entitled to monthly social security cash benefits or payments from the railroad retirement system are eligible for benefits under the HI program. Also, disabled persons entitled to cash benefits under the social security or railroad retirement programs are eligible for HI benefits. A person must be disabled for 5 calendar months and then entitled to 24 months of cash benefits before becoming eligible for HI benefits. Thus, Medicare coverage begins the 30th month after the first full calendar month of disability. HI protection also extends to persons who have ESRD and require renal dialysis or a kidney transplant if they are currently insured, entitled to monthly social security benefits, or are the spouses or dependent children of such insured persons.

Persons entitled to benefits under the HI program and most other persons 65 years of age or over may voluntarily enroll in SMI. Persons may terminate SMI enrollment by not paying premiums. Under the State buy-in system, a State government may enroll and pay SMI premiums for eligible aged and disabled individuals who are also covered by the Medicaid program.

As of July 1, 1991, it was estimated that 97% of the aged residing in the United States were enrolled in HI and/or SMI. Among the aged population not covered by Medicare are federal employees who retired prior to 1983 and never having worked under social security. Those who never worked enough work credits (quarters of coverage), such as migrant workers and members of religious orders, are also not eligible for Medicare. However, most persons 65 years of age or over who are ineligible for HI coverage are permitted to enroll voluntarily by paying a monthly premium. To obtain premium-HI, the enrollee must also obtain SMI coverage. Also not represented in the Medicare population are many persons who continue to work beyond age 65 and do not apply for Medicare. These late retirees do not receive Social Security cash benefits since their income exceeds the maximum amount allowed for retirement benefits. Some are covered by employer's health insurance plans and fail to apply for HI coverage.

HCFA maintains the Health Insurance Master File (HIM) which identifies each person entitled to Medicare benefits. Identification of each record is based on a claim number which consists of a person's Social Security or Railroad Retirement Board number and a one or two position beneficiary identification code (BIC). The BIC portion of the claim number describes the type of benefit that entitles individuals to Medicare coverage. The HIM file is updated daily with current maintenance and utilization information. To tabulate Medicare enrollment data, a skeletonized version of the HIM file, known as the Health Insurance Skeleton Eligibility Write-off (HISKEW), is produced quarterly from the HIM file.

The HISKEW file provides a fixed frame of reference and was used to create a frame to select the beneficiaries. Specifically, beneficiaries were selected according to the following criteria: (1) Their health insurance claim numbers ended in 05, 20, 45, 70, or 95 (this is the standard 5% sample studied in HCFA); (2) They were entitled to Medicare Part A and/or Part B benefits on or before January 1, 1991; (3) They were alive on the selection date; (4) They lived in one of the sample ZIP fragments.

The sample ZIP fragments were selected from a county-by-ZIP summary of the 1988 CMHS file prepared by HCFA. The CMHS is a micro file containing beneficiary utilization of all Medicare benefits for the same 5% sample of beneficiaries as the HISKEW.

5. Coverage Ratios and Design Effects

Table 1 shows two estimates of the reference population. The first was obtained by tabulating all beneficiaries of record as of January 1, 1991, according to the 5% March 1991 HISKEW. The second was obtained by tabulating the baseweights of the MCBS sample. The third column gives the MCBS sample estimate as a percentage of the HISKEW tabulation. The last column shows the standard error of the undercoverage rate. As is evident, there is an undercoverage of 2.3% in the sample. The undercoverage varies by age, sex and region. The difference between the HISKEW and the MCBS are all statistically significant at the 5% level, except for the age group 0-44.

Table 1. Comparison of sample and administrative estimates.

	HISKEW[2]	MCBS	Covered	Standard Error
0-44	1,052,560	1,050,534	99.8%	0.18%
45-64	1,967,060	1,953,398	99.3%	0.33%
65-69	7,158,880	6,936,647	96.9%	0.49%
70-74	8,611,460	8,352,464	97.0%	0.49%
75-79	6,638,420	6,517,285	98.2%	0.38%
80-84	4,543,940	4,439,417	97.7%	0.42%
85+	4,233,060	4,157,516	98.2%	0.35%
Total	34,205,380	33,407,261	97.7%	0.17%
Male	14,461,700	14,027,976	97.0%	0.30%
Female	19,743,680	19,379,285	98.2%	0.20%
Northeast	7,491,440	7,458,769	99.6%	0.16%
South	11,641,560	11,434,088	98.2%	0.31%
Midwest	8,461,820	8,006,415	94.6%	0.44%
West	6,178,620	6,062,656	98.1%	0.36%
Puerto Rico	431,940	445,332	103.1%	

Our investigation determined that most of the undercoverage was due to missed ZIP fragments. These missed ZIP fragments occurred on the March 1991 HISKEW but not on the 1988 CMHS. New ZIP and county coding errors are constantly being committed, and the Post Office periodically realigns ZIP code boundaries and creates new ZIP codes. Furthermore, only the 5% CMHS sample was used in sampling. There were undoubtedly small ZIP fragments that would have only appeared in the full beneficiary list. As beneficiaries move, however, some of these rare ZIP fragments may enter the 5% sample. The basic problem was that a 1988 listing of beneficiaries was used to assign measures of size to ZIP fragments. Thus persons living in new ZIP fragments (created either by the Post Office or by human error) and persons moving into ZIP fragments that did not appear in the 1988 5% sample list had no chance of selection for the initial round of the MCBS in 1991. These missed ZIP fragments accounted for 1.7 of the missing 2.3 points. It is not known why the remaining 0.6 points were missed, but at least some of it is due to the persons with invalid county codes and invalid ZIP codes that we deliberately excluded from sampling. The plan for MCBS is to draw a supplementary sample of beneficiaries for 1993 to compensate for this missed population.

A number of other possible reasons were examined and rejected. One possibility concerned beneficiaries with foreign addresses. Inclusion of these in the HISKEW tabulations could have caused the appearance of undercoverage in MCBS. However, it was verified that they were not included. Another possibility was some omission of county codes. In the ZIP code sampling, every county code was checked to make sure that it was legitimate. Those that were not legitimate we replaced with imputed legitimate county codes. All of this was done on the CMHS summary file rather than on a HISKEW summary file, but it seems unlikely that there are many county-coding errors that are unique to the HISKEW. The one difference in county coding that was detected between the two files (blank versus triple-zero county codes) was corrected for in the sampling and weighting. A third possibility that occurred to us was some error in the weighting. Our weighting procedures had been reviewed. If such an error exists, it eluded us.

The undercoverage of 2.3% is not a serious problem. Undercoverage rates are typically higher in area samples. The Census Bureau currently has a 5% undercoverage rate in its carefully conducted household samples. The National Household Interview Survey (NHIS) has an undercoverage rate of 8% for the 65 and older population. Furthermore, ratio adjustment will have reduced any geographic bias that the undercoverage causes. Despite the small size of the undercoverage, we are fielding a coverage improvement sample in the missed ZIP fragments in the fall of 1992.

Table 2 shows design effects for selected statistics. The design effect is the ratio of the variance of the estimate obtained from the MCBS sample to the variance of the estimate obtained from a simple random sample of the same sample size. The estimated design effects shown in the table are themselves subject to sampling errors. This is clearly evident in the design effect for 70-74 year olds with income below the median. It is not reasonable for this design effect to be 1.03 while the design effects for neighboring age brackets are 1.72 and 1.75. Some kind of smoothing is therefore needed before using the design effects. (Smoothing may be one of the topics of a future report.) As would be expected on theoretical grounds, design effects are larger for broad domains than for narrow domains. This is partly due to larger cluster sizes for the broader domains and to differential sampling rates across the age domains. It has not been estimated how much is due to the clustering at the county level, how much to the clustering at the ZIP code level, and how much to the differential sampling rates. Also predictable is that the effects of clustering are more evident in the socio-economic variables than in the health status variable.

Table 2. Selected Design Effects.

Characteristic	AGE									
	0-44	45-64	65-69	70-74	75-79	80-84	85+	Total	Males	Females
Hypertension	0.92	1.14	0.98	1.07	0.96	1.19	1.23	1.22	1.42	1.24
Medicaid Participation	1.22	1.03	0.96	0.99	1.41	1.35	1.34	1.82	1.18	1.53
Below Median Income	1.15	1.36	1.72	1.03	1.75	1.19	1.21	3.36	2.16	2.40

6. Address Quality and Mobility

Initial contact with an MCBS sample person (SP) was made by means of a respondent letter. The letter served to legitimize the study to the SP by providing a brief description of the survey with a particular emphasis on the survey sponsorship by a federal government agency (HCFA). The letter was also intended to prepare the SP for the in-person visit by an interviewer. The general thinking was that this mostly elderly population was more likely to be suspicious and reluctant to allow strangers into their homes to conduct an interview than the general population. In order to insure that the letters had the most accurate address available, all addresses provided by HCFA underwent a cleaning process. The first stage of this process was an automated address cleaning. At this stage, each of the address fields was looked at for recognizability. Any addresses containing a field which did not meet the program algorithm were classified as problems. Of the 15,215 original sample addresses, 1084 or 7% were classified as having a problem.

All of the addresses identified as problems then underwent a manual review and cleaning process. The review found that the majority of problematic addresses seem to be a function of data entry errors. The most common problems were SP names entered in address fields, ZIP codes which did not exist, and ZIP codes which did not match the city or state in the address field. The manual review process was able to correct all of the problematic addresses to the point where a letter could be mailed to the SP with confidence it could be delivered.

Letters were mailed to all but six of the SPs who were living out of the country and therefore ineligible to participate in the survey. These six cases were identified in the manual review process as having an out of country address, however they had an incorrect ZIP code which caused them to be included in the sample.

Letters to the SPs were mailed with an "Address Correction Requested" preprinted on the envelope. Of the 15,209 letters mailed, 14,511 or 95.5% were delivered as addressed and 322 or 2.1% were delivered with a change of address returned to Westat. Three hundred and seventy six or 2.5% of the letters were unable to be delivered as addressed. The reasons given by the post office for the undeliverable letters can be seen below in Table 3.

Table 3. Reasons For Undeliverable Advance Letters.

REASON	Number	Percent
Attempted - Not Known	77	20
Reason Unknown	73	19
Insufficient Address	69	18
No Such Number	36	10
Unable to Forward	34	9
Forwarding Order Expired	32	9
Addressee Unknown	23	6
Moved, Left No Address	18	5
No Such Street	11	3
Box Closed - No Order	3	1
Total Not Delivered	376	100

The reasons given by the post office for not being able to deliver these letters for the most part seem to be explainable by random data entry error. However, one category, "Forwarding Order Expired" seems to be a curiosity. In order for a letter to fall into this category the address on the advance letter would need to have been a valid address for the SP at some point in time. Furthermore, the SP would have needed to file a change of address with the post office over a year prior to the date of attempted delivery in order for the letter to be classified as "Forwarding Order Expired" by the post office. This means that the address being carried on the HCFA list was over a year out of date for these 32 SPs.

The 376 cases with bad addresses procedurally fell into a field tracing status. The field interviewers began a tracing protocol which included: visiting the last known address, asking neighbors for locating information and contacting local governmental agencies such as the motor vehicle department and the post office to obtain updated address information. As a result of the resourcefulness of the interviewers, only 184 cases (1.2% of the sample) were finalized as unlocatable at the end of the field period.

It should be noted that these 184 cases are not just a subset of the original 376 cases with bad addresses, but also included a number of cases with a valid address to which mail for the SP could be delivered but the SP could not physically be located. The most typical situation was an SP who had a post office box but would not respond to any request for an interview. In situations like this, the field interviewer would leave a letter asking the SP to call the toll free Westat respondent line, the HCFA respondent hot line or to contact the interviewer directly. Only after repeated attempts were made to prompt the SP to respond were these cases finalized as unlocatable.

One other type of case which was finalized as unlocatable should be noted. These are cases in which the SP was homeless, and had a place where mail was delivered, but the interviewer was unable to locate the SP in person. Most typically the mail would go to a shelter where the person was known, and the shelter would then hold the mail for the SP to pick up. Because the SPs would only sporadically stop by to pick up mail, it was extremely difficult for the interviewers to contact these types of SPs. Once again, after repeated attempts were made to contact the SP, this type of case was classified as unlocatable. As a whole it seemed that the unlocatable cases constituted a pool of SPs who did not want to be found for a variety of reasons, whether that reason was just the desire of the SP to maintain anonymity and not participate in the survey or whether the reason was a function of the transient nature of the residence such as the homeless.

One other subset of non-response which needs to be looked at when trying to understand address quality are the cases which ended up being finalized as "Out of area". While only 64 cases (0.4% of the sample) ended up with this final status, it points to an anomaly in the addresses. The category itself was created because of operational considerations and was defined to comprise SPs who were residing over 150 miles from a sample PSU, and thus from an interviewer. It was clear that interviewing SPs this far away from the PSU was prohibitively expensive. Although some of the SPs residing outside the PSU moved after the initial addresses were selected, the majority were a result of the mailing address being different than the place of residence. In some cases mail was delivered in care of a relative, attorney, or bank, and forwarded to the SP. In other cases, the SP maintained a post office box in the PSU but lived outside of the PSU. The most interesting of these were SPs living along the border with Mexico, who maintained post office boxes in border towns in the United States, while they actually lived across the border in Mexico. SPs who lived within 150 miles of a sample PSU were routinely interviewed. Although we do not have an accurate count of SPs who live within 150 miles of a sample PSU we estimate that we are interviewing 300 to 500 SPs (2 to 3% of the sample) who reside outside of the PSU boundaries. The vast majority of these were not a function of a move after the initial sample was drawn, but rather a function of the HCFA address not describing the place of residence.

7. MCBS Response Rates

The MCBS is actually comprised of two separate surveys, each with very different operational procedures and each yielding very different response rates. The MCBS sample consists of SPs residing in ordinary households, who were defined as being in the Community Component and SPs residing in institutional settings and receiving long term care who were defined as being in the Facility Component. The contact procedures for the two components are completely different. For the SPs defined as belonging in the Community Component, each contact is with the individual SP or the SP's designated proxy. In the Facility Component, the SP is never directly contacted. The initial letter of introduction described in the previous section is mailed directly to the facility administrator. The administrator is then contacted by an interviewer and asked to designate a staff member who can answer questions about the SP. The response rate for the Facility Component consequently is a function of facility cooperation instead of SP cooperation.

Of the 15,411 individuals sampled, 881 were deemed ineligible (deaths prior to being interviewed), leaving us with a working sample of 14,530 SPs. Of these, 13,541 or 93.2% were classified as being in the Community Component and the remaining 989 or 6.8% were classified as being in the Facility Component. The final status of cases in each of these components can be seen in Table 4.

Table 4. Final Status of Eligible Sample.

	Communtity Component		Facility Component		Total Eligible Sample	
Final Status	Number	Percent	Number	Percent	Number	Percent
Complete	11,735	86.7	942	95.2	12,677	87.2
Refusal	1,376	10.2	22	2.2	1,398	9.6
Unlocatable	180	1.3	4	0.4	184	1.3
Unavailable	73	0.6	3	0.3	76	0.5
Out of Area	57	0.4	7	0.7	64	0.4
Incompetent, No Proxy available	53	0.4	0	0.0	53	0.4
Other Non-Response	32	0.2	10	1.0	42	0.3
Breakoff	29	0.2	1	0.1	30	0.2
Language Problem	6	0.0	0	0.0	6	0.0
TOTAL	13,541	100.0	989	100.0	14,530	100.0

There was a substantive difference between the response rates for the two components, 86.7% for the Community Component as compared to 95.2% for the Facility Component. This magnitude of difference can also be seen in the refusal rates which were 10.2% for the Community Component and 2.2% for the Facility Component. The Community response rate was well within our expectations based on the 1987 National Medical Expenditure Household Component, a combined screener and round 1 response rate of 85%.

It should be noted that the 85% NMES response rate is for the general population and we would have expected a somewhat lower response rate with the elderly. The response rate for the Facility Component was also within our expectations based on the Institutionalized Population Component (IPC) of NMES. The IPC response rate for the facility questionnaire was 94.9%.

This study also modeled response rates using a logistic regression model and detailed predictor variables from administrative files. The analyses and findings will be presented in a future paper.

[1] A supplement of 196 beneficiaries in the 85+ age group was later added to the original sample to be interviewed. This supplementary sample was to compensate for the "smaller that desired" original sample size. The designated sample size for the 85+ age group in the original sample had been determined using a death rate that was too low.

[2] March 1991 5% HISKEW, weighted up by 20. Excludes foreign addresses, deaths prior to 1991 and post-1990 new eligibles.