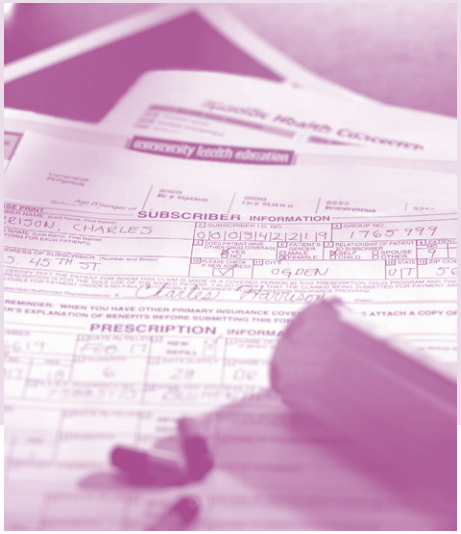

TECHNICAL DOCUMENTATION FOR THE MEDICARE CURRENT BENEFICIARY SURVEY

Overview

The Medicare Current Beneficiary Survey (MCBS) is a continuous, multipurpose survey of a nationally representative sample of aged and disabled Medicare beneficiaries sponsored by the Centers for Medicare and Medicaid Services (CMS). In 2007, the initial sample included approximately 17,177 beneficiaries residing in households and long-term care facilities.[1] The survey provides comprehensive data on health and functional status, health care expenditures, and health insurance for Medicare beneficiaries. A key feature of the survey is its longitudinal design. Currently, each sample person is interviewed 3 times a year over 4 years, regardless of whether he or she resides in the community or a facility, or transitions between community and facility settings. (For a description of the MCBS, see G.S. Adler, Summer 1994, A Profile of the Medicare Current Beneficiary Survey, Health Care Financing Review, 15(4): 153-163.)

Sample Design

The target population consists of aged and disabled beneficiaries enrolled in Medicare Part A (hospital insurance), or Part B (medical insurance), or both, and residing in households or long-term care facilities in the United States and Puerto Rico. Sample persons are selected from Medicare enrollment files to be representative of the Medicare population as a whole and the following age groups: under 45, 45 to 64, 65 to 69, 70 to 74, 75 to 79, 80 to 84, and 85 and over. To ensure that annual samples yield enough persons with long-term care facility stays to produce statistically reliable data, disabled persons under age 65 and very old persons age 80 and over are oversampled.

The MCBS was originally designed as a longitudinal survey in which Medicare beneficiaries would be followed indefinitely. Its initial sample (the 1991 panel) was selected by using a stratified, multistage area probability design. Three stages of selection were used in sampling beneficiaries. The first stage was to select a nationally representative stratified sample of 107 primary sampling units (PSUs) consisting of metropolitan statistical areas or clusters of nonmetropolitan counties. The second stage was to select ZIP code clusters within sample PSUs. The third stage consisted of selecting beneficiaries within the sampled ZIP code clusters.

In 1992 and 1993, the 1991 panel was supplemented during the September-December interview period to compensate for sample attrition (i.e., deaths, disenrollments, and refusals) and to represent newly enrolled beneficiaries. However, in 1994, approximately one-third of the sample was rotated out of the MCBS after the round 12 interviews, and replaced by a supplemental sample of the same size. The change in supplemental sampling reflects a decision to shift from a longitudinal survey to a rotating panel design. In the rotating panel design chosen for MCBS, four overlapping panels of Medicare beneficiaries will be surveyed each year. Each panel contains a nationally representative sample of beneficiaries who will be interviewed 12 times to collect 3 complete years of utilization data. All four panels are included in the Access to Care files, while only three panels are used in the Cost and Use files, since the panel that is being retired during a calendar year is not asked about medical utilization for that year.

Survey Operations

Field work on the MCBS is conducted for CMS's Office of Strategic Planning by Westat, a survey research firm with offices in Rockville, Maryland. Data collection for Round 1 began in September 1991 and was completed in December 1991. Subsequent rounds of data collection, which involve reinterviewing the same sample persons (or their proxies—see below), begin every 4 months. Interviews are conducted regardless of whether the sample person resides at home or in a long-term care facility, using the version of the questionnaire appropriate to the setting.

[1] Beneficiaries living in households are referred to as community residents in this sourcebook.

In 2007, data were collected from 11,995 beneficiaries for the Cost and Use file. The final sample included 10,977 persons who lived in the community for the entire year, 807 persons who lived in long-term care facilities for the entire year, and 211 persons who lived part of the year in a community and part of the year in a long-term care facility. Interview strategies and survey instruments used to collect data are described below.

Repeat Interviews. The MCBS is a longitudinal panel survey, with sample persons interviewed 3 times a year over 4 years to form a continuous profile of their health care experience. The design allows MCBS data users to track change in insurance coverage and other personal circumstances. For example, users can observe processes such as persons moving from their homes to long-term care facilities, or persons in communities spending down their assets on health care.

The Community Interview. Sample persons in the community are interviewed through computer-assisted personal interviewing (CAPI) survey instruments. The CAPI program automatically guides the interviewer through questions, records the answers, and compares beneficiary responses to edit specifications for accuracy and relationships to other responses. CAPI improves data collection and lessens the need for after-the-fact editing and corrections. It guides the interviewer through complex skip patterns and inserts followup questions where key data are missing from the previous round. When the interview is completed, CAPI allows the interviewer to transmit the data by telephone to the home office computer.

The interviews yield a time series of data on utilization of health services, medical care expenditures, health insurance coverage, sources of payment for health services, health status and functioning, and beneficiary information such as income, assets, living arrangement, family assistance, and quality of life. To improve the accuracy of the data, respondents are requested to record medical events on calendars provided by the interviewer, and they are also asked to save Explanation of Benefit forms from Medicare, as well as receipts and statements from private health insurers. To assist in reporting data on prescription medicines, respondents are asked to bring to the interview bottles, tubes, and prescription bags provided by the pharmacy.

An effort is made to interview each sample person directly. However, each sample person is asked to designate a proxy, usually a family member or close acquaintance, in case he or she is physically or mentally unable to do the interview. On average, about 12 percent of the community interviews in each round are conducted by proxy. The following instruments are used in community interviews:

- ***The Baseline Questionnaire:*** Collects health insurance, household composition, health status, access to and satisfaction with medical care, and demographic and socioeconomic information for supplemental sample beneficiaries living in household units in the community. Selected information from this questionnaire—primarily health status, and access to and satisfaction with care—is updated annually for continuing sample persons living in the community using The Community Supplement to the Core Questionnaire.

- ***The Community Core Questionnaire:*** Collects detailed health insurance, medical care use, and charge and payment information, and updates household composition. This questionnaire is asked in every round except the initial one. Additional supplemental questions are added to the core questionnaire in selected rounds to gather information about specific topics, including detailed information about the sample person's income and assets in the spring-summer round of data collection.

The Facility Interview. MCBS data collectors in long-term care facilities use a similar but shortened version of the community instrument. A long-term care facility is defined as having three or

more beds and providing long-term care services throughout the facility or in a separately identifiable unit. Types of facilities participating in the survey include nursing homes, domiciliary or personal care facilities, distinct long-term care units in a hospital complex, mental health facilities and centers, assisted living and foster care homes, and institutions for the mentally retarded and developmentally disabled.

If an institutionalized person returns to the community, a community interview is conducted. If he or she spends part of the reference period in the community and part in an institution, a separate interview is conducted for each period of time. Hence, a beneficiary can be followed in and out of facilities, and a continuous record is maintained regardless of where the person resides.

Because long-term care facility residents often are in poor health and many facility administrators prefer that patients not be disturbed, the survey collects information about institutionalized patients from proxy respondents affiliated with the facility. Nurses or other primary care givers usually respond to questions about physical functioning and medical treatment of the sample person. Billing office workers usually respond to questions about charges and payments.

The survey instruments used to collect data for persons in long-term care facilities were converted to CAPI in 1997. The following instruments are used in facility stay interviews:

- ***The Facility Screener:*** Collects information on facility characteristics such as type of facility, size, and ownership. It is used during the initial interview, and in each fall round thereafter.

- ***The Baseline Questionnaire:*** Collects information on health status, insurance coverage, residence history, and demographics for supplemental sample beneficiaries in facilities and new admissions from the continuing sample. Selected information from this questionnaire—primarily health status—is updated annually for continuing sample persons residing in facilities using an abbreviated version, The Facility Supplement to the Core Questionnaire.

- ***The Facility Core Questionnaire:*** Collects facility use data, and charge and payment information. This questionnaire is asked in every round except the initial one.

The conversion of the facility instruments to the CAPI version caused certain disruptions in the trend data for full-year facility residents, because some questions/items are phrased differently in the CAPI version from those in the Paper-and-Pencil version. Variables in the Health Segment affected the most include self-reported health status, functional limitations, and most of the diseases/conditions presented in data tables in Section 2 of Chapter 3. Therefore, caution needs to be exercised in examining the health trend data for full-year facility residents presented in this series of sourcebooks.

MCBS PUBLIC USE FILES

To date, CMS has released public use files (PUFs) on access to care for calendar years 1991 through 2009, and on cost and use for calendar years 1992 through 2008.

Access to Care

The Access to Care PUFs provide "snapshot" estimates of the characteristics of the Medicare population who were enrolled on January 1 and were still alive and eligible for the survey in the fall of each year. They contain information on access to and satisfaction with care, health status and functioning, and demographic and economic characteristics of the sample population. Access to Care PUFs

also contain summarized utilization and program payment data from Medicare claims, but they do not include survey-reported information on health care use and expenditures. By omitting the survey-reported information, these PUFs can be produced quicker than cost and use files, which contain complete information on the cost and use of health care services.

Cost and Use

The 2007 Cost and Use file is the sixteenth in an annual series of files containing comprehensive data on the cost and use of medical services by the Medicare population.[2] It links Medicare claims to survey-reported events, and provides complete expenditure and source of payment data on all health care services, including those not covered by Medicare. Expenditure data were developed through a reconciliation process that combines information from survey respondents and Medicare administrative files. The process produces a comprehensive picture of health services received, amounts paid, and sources of payment. The file can support a broader range of research and policy analyses on the Medicare population than would be possible using either survey data or administrative claims data alone.

The strength of the file stems from the integration of information that can be obtained only from a beneficiary, and Medicare claims data on provider services and covered charges. Survey-reported data include information on the use and cost of all types of medical services, as well as information on supplementary health insurance, living arrangements, income, health status, and physical functioning. Medicare claims data include use and cost information on inpatient hospitalizations, outpatient hospital care, physician services, home health care, durable medical equipment, skilled nursing home services, hospice care, and other medical services.

File Structure

The Cost and Use file contains information on nine types of services: dental, facility stays, institutional utilization, inpatient hospital stays, outpatient hospital care, physician/supplier services, hospice care, home health care, and prescription drugs. As an aid to file users, the data have been provided at the event-level, the type-of-service level, and the person-level. The hierarchical structure allows analysts to use the appropriate file level for their research, avoiding the need to process all the detailed event records in the file. For example, differences in per capita health spending between men and women can be analyzed directly from person-level summary records. Similarly, differences in hospital stays by race can be analyzed directly from type-of-service summary records. Event-level records would be used for more detailed analyses; e.g., comparisons of average length of long-term facility stays or average reimbursements per prescription drug. The content of each level of data is briefly described below.

Event-level data. The event-level data consist of separate files for each of the nine event types in the Cost and Use file, except hospice care and home health care. For each event in a file, cost and sources of payment are shown. Charge and payment data have been edited and imputed, if necessary, to make a complete payment picture for each event. Hospice care and home health care are not shown at the event-level because these two service categories were created from Medicare claims data at the type-of-service level. There are a total of 927,051 records in the seven event-level files.

Type-of-service summary data. The type-of-service summary file includes a record for each of the nine service categories in the Cost and Use file. The file contains a summary of all payers, costs, and use for each sample person at the type-of-service level, for a total of 107,955 records. Within each type-of-service record, separate payer amounts are shown for the 11 payer categories in the Cost and Use file. Payer totals are shown two ways: as the sum of event-level pay-

[2] Detailed documentation of the CY 2007 Cost and Use file is available from the Centers for Medicare and Medicaid Services, Office of Research, Development, and Information, in Baltimore, Maryland.

ments and in adjusted form. Adjusted payments are necessary because some sample persons had gaps in their coverage (e.g., a respondent missed an interview during the year). To account for information that was not reported for the gap periods, payer amounts were adjusted for differences in Medicare-covered days and days covered by the interview reference periods. Most of the adjustments were for services not covered by Medicare, since CMS's administrative files have claims for covered services provided to fee-for-service beneficiaries during gap periods.

Person-level summary data. The person-level summary file has one record for each of the 11,995 sample persons in the 2007 Cost and Use file. Payments by source have been summarized across service categories to show one total for each type of service and one total for each source of payment. Again, payment amounts are shown as totals from the event-level files and in adjusted form. This sourcebook uses the adjusted amounts.

The Sample

The original MCBS sample included Medicare beneficiaries who resided in the United States or Puerto Rico on January 1, 1991, and who were enrolled in one or both parts of Medicare at the time of their Round 1 interview. Round 1 was fielded from September through December of 1991. Except for a small number of individuals who died or whose coverage terminated subsequent to their interview, the overwhelming component of this group was the "always-enrolled" 1991 population. This group consisted of persons who had enrolled in Medicare by January 1, 1991, and were still covered by Medicare on December 31, 1991. Selected data on the Round 1 always-enrolled sample were released as the CY 1991 Access to Care file.

The always-enrolled concept also was used to determine the sample populations in the Access to Care releases in subsequent years. Official Medicare program statistics, however, usually cover all persons entitled to Medicare during the year, including those entitled for all or part of the year, as well as beneficiaries who died during the year. This mix of continuing enrollees, accretions, and terminations is referred to as the "ever-enrolled" population, or everyone who was enrolled in Medicare for any period during the year.

Special steps are taken to expand sample coverage in the Cost and Use files to include all beneficiaries who were ever enrolled during the calendar year. The steps are necessary because Cost and Use files will be used to analyze total and per capita expenditures on health care by the entire Medicare population. Omitting part-year enrollees and persons who died during the year could substantially bias the results of these analyses.

To develop the ever-enrolled population in 2007, supplemental samples were used to add part-year beneficiaries to the Cost and Use file. A supplemental sample is drawn each year to account for growth in the Medicare population and to replace survey persons who died or left the survey during the previous year. Sample replenishment is used primarily to ensure that each calendar year file adequately represents the entire Medicare population, but it also can be used to identify new sample persons who were covered by Medicare in the sample year but were missing from the original sampling list. Beneficiaries from supplemental samples in Rounds 49 and 52, who enrolled during 2006 or 2007, were added to the samples from Rounds 40, 43, and 46 to create an ever-enrolled population for calendar year 2007.

The 2007 Cost and Use file, therefore, consists of a composite of persons who were (1) continuously enrolled from January 1, 2006; (2) newly enrolled in 2006; or, (3) newly enrolled in 2007. The number of persons in each group is shown in Table A-1, where newly enrolled beneficiaries after 1992 are referred to as "accretes." The pre-2006 accretes represent persons who were enrolled in Medicare before 2006 and still living in 2007.

Table A-1 2007 Cost and Use File Sample

Sample Status	Number of Persons
Pre-2006 Accretes (Panels 2004, 2005, & 2006)	11,361
2006 Accretes (Panel 2007)	335
2007 Accretes (Panel 2008)	299
Total	11,995

Newly enrolled sample persons from Rounds 49 and 52 are colloquially referred to as "ghosts" because they did not become eligible for Medicare in time to be selected as part of the sample that received all three 2007 interviews. Thus the sample persons who represent 2006 and 2007 accretes (i.e., beneficiaries who were newly enrolled in Medicare in 2006 or 2007) have incomplete or missing survey data for 2007.

Utilization data for ghosts are included in the 2007 Cost and Use file at the type-of-service and person summary levels, even though they were not interviewed until late 2007 (Round 49) if they were new Medicare enrollees in late 2006, or late 2008 (Round 52) if they were new Medicare enrollees in 2007. While survey data on service use and costs were not available for ghosts, complete profiles of Medicare-covered service use by fee-for-service ghosts were available from administrative bill files. To estimate total service use and costs for the entire sample, ghosts were matched to donor beneficiaries in the 2007 file based on common Medicare use profiles. The donor records were used to impute noncovered services for fee-for-service ghosts and all services for Medicare risk HMO ghosts.[3] This imputation process provided estimates of missing cost and use data for the ever-enrolled population in the 2007 Cost and Use summary files.

Access to Care or Cost and Use Data?

The Cost and Use file is more comprehensive than the previously released Access to Care files because it contains the always-enrolled population, as well as persons entering or leaving the Medicare program during the year. The latter group of beneficiaries is essential in producing accurate estimates of total expenditures because it includes beneficiaries who died during the year. Tabulations of Medicare claims for the MCBS sample, for example, show that persons who died in the year represent less than 5 percent of the Medicare population, but they account for more than 15 percent of Medicare payments. On average, persons who died during the year have spending levels over 4 times higher than persons continuously enrolled for the entire year.

Another difference between the two files relates to the reporting of expenditures on health care. The Access to Care files contain only Medicare-covered service data, even though Medicare has been previously estimated to cover less than one-half of the overall care expenses of its enrollees (D.R. Waldo, S.T. Sonnefeld, D.R. McKusick, et al., Summer 1989, "Health Expenditures by Age Group, 1977 and 1987," Health Care Financing Review, 10(4): 111-120). The Cost and Use file, in contrast, includes expenditures on all health care services, whether or not they are covered by Medicare. Two significant expenditure categories not covered by Medicare are prescription drugs and long-term facility care.

Users whose analyses require the entire Medicare population or all health care services should use the Cost and Use files rather than the Access to Care files. Users who are interested in the continuously enrolled Medicare population or Medicare-covered services only may prefer to use the Access to Care files. In addition, the latter set of files can be used for some types of longitudinal analyses, such as a comparison of change in health status from year to year.

[3] Medicare risk HMO contractors do not submit claims to Medicare. As a result, Medicare does not have a record of covered or noncovered services provided to beneficiaries in these plans.

Users are cautioned against mixing data from the two types of files to estimate change over time. For example, 2007 Cost and Use file data on health status should not be compared to 2007 Access to Care file information since the results will be confounded by differences in the two populations. Unless the two files are subset to a common set of sample persons and appropriate weights are assigned, it would be difficult, if not impossible, to determine whether health status had changed over time.

Response Rates and Missing Data

The sample for the 2007 Cost and Use file originally contained 4,968 beneficiaries from Round 40; 5,367 beneficiaries from Round 43; 5,880 beneficiaries from Round 46; 474 beneficiaries from Round 49, who became eligible for Medicare in 2006; and 488 beneficiaries from Round 52 who became eligible for Medicare in 2007. The beneficiaries from Rounds 40, 43, and 46 all survived until 2007. The overall response rate was 69.8 percent for a final sample of 11,995 persons. Response rates are shown in Table A-2.

Table A-2 2007 Cost and Use File Sample Response Rates

Panel	Sample Size	Respondents	Response Rate
Round 40	4,968	3,319	66.9%
Round 43	5,367	3,702	69.0%
Round 46	5,880	4,340	73.8%
Round 49	474	335	70.7%
Round 52	488	299	61.3%
All	17,177	11,995	69.8%

As in any survey, some respondents did not supply answers to all questions. Item nonresponse rates are low in the 2007 Cost and Use file, but analysts still should be aware of missing data. For example, the number of missing responses and item nonresponse rates for several variables are shown in Table A-3.

Table A-3 2007 Item Nonresponse for Selected Variables

Variable	Missing	Percentage of Total
Race/Ethnicity	30	0.3%
Education	209	1.7%
Marital Status	33	0.3%
Gender	0	0.0%
Age	0	0.0%
General Health	94	0.7%

Since data for most variables are fairly complete, imputations were kept to a minimum in the 2007 Cost and Use file. Each user can decide how to handle missing data. A simple approach is to delete records with missing data, but the cumulative effect of deleting each record with missing data can significantly reduce the data available for analysis. Other approaches would be to create an "unknown" or "missing" category within each variable distribution or to assume the distribution of missing data is the same as that of reported data. The latter approach was often used in creating tables for this sourcebook.

Another alternative for handling cases with missing data is to impute the missing values. This approach was used to create complete information on beneficiary income and expenditures for health care in the Cost and Use file. Imputations were performed on these variables because income and expenditure data are key elements of the file. In imputing the expenditure data, all partial information from survey respondents was preserved to the extent possible, and health insurance data from the survey and Medicare administrative files were used to identify potential payers. Analytic edits and hot-decking methods were used to estimate missing payments and charges.[4]

[4] The technical appendixes in the 2007 Cost and Use file documentation detail the imputation methods used to complete the expenditure data.

COST AND USE FILE STATISTICS

The 2007 Cost and Use file contains a cross-sectional weight for each of the 11,995 beneficiaries in the data set. These weights reflect the overall selection probability of each sample person and include adjustment for survey nonresponse and post-stratification to control totals based on accretion status, age, sex, race, region, and metropolitan area status. The weights inflate the sample to the ever-enrolled Medicare population in 2007, and were used in producing all tables in this sourcebook. In general, the weights should be used to estimate population totals, percentages, means, and ratios.

Sampling Error

Sampling error refers to the expected squared difference between a population value (a parameter) and an estimate derived from a sample of the population (a statistic).[5] Because the MCBS is a sample of Medicare beneficiaries, statistics derived from the sample data are subject to sampling error. The error reflects chance differences between estimates of a population parameter that would be derived from different samples of the Medicare population. Nearly any MCBS estimate of a population parameter (e.g., a percentage, mean, ratio, or count of persons or events) would be affected by the sampling error.

Standard errors have been calculated for all statistics reported in the detailed tables in this sourcebook in order to assess the impact of sampling variability on the accuracy of the estimates. Data from Table 1.1 of this sourcebook, for example, indicate that 44.17 percent of all Medicare beneficiaries are between the age of 65 and 74. The standard error of this estimate (0.41 percent) can be used to assess its statistical reliability by constructing a confidence interval that would contain the true value of the population parameter with some given level of confidence.

The confidence interval can be viewed as a measure of the precision of the estimate derived from sample data. For example, an approximate 95 percent confidence interval for statistics in this sourcebook can be calculated by using the formula

$$\pi = P \pm 1.96 \text{ x } (\text{estimated standard error}),$$

where ¼ is the unknown population proportion and P is the calculated (weighted) sample proportion. Based on this formula, the approximate 95 percent confidence interval for the estimated proportion of Medicare beneficiaries between the age of 65 and 74 is 44.17 percent plus or minus 0.80 percent. This is a relatively "tight" confidence interval, suggesting that the MCBS data provide a reliable estimate of the true proportion of beneficiaries between the age of 65 and 74. The chances are about 95 in 100 that the true population proportion falls between 43.37 percent and 44.97 percent.

Another measure of statistical reliability is the relative standard error (RSE) of an estimate. The RSE of an estimate x is calculated by dividing the standard error of the estimate, SE(x), by the estimate, and expressing the quantity as a percent of the estimate, i.e.,

$$RSE = 100\left(\frac{SE(x)}{x}\right).$$

Using data from the previous example, the RSE of the estimated proportion of Medicare beneficiaries between the age of 65 and 74 is 0.93 percent (100 x (0.41/44.17)). An RSE of less than 10 percent would suggest that the estimate is statistically reliable. Statistical reliability of an estimate decreases as the RSE increases.

Many of the statistics in this sourcebook are presented by subgroup, some of which are based on relatively small sample sizes. Estimates for these small subgroups can be subject to very large sampling errors. Therefore, it may be desirable in some instances to combine such sub-

[5] This discussion ignores errors caused by factors such as imperfect selection; bias in response or estimation; and errors in observation, measurement, or recording.

groups with a similar group for analysis purposes. For example, if X_s is an estimated total for the small subgroup, and Xt is the corresponding estimate for the group with which it is combined, then the combined estimate, X_c, is given by $X_c = X_s + X_t$, and the standard error of the combined estimate (SE(Xc)) can be approximated as

$$SE(X_c) = \sqrt{[SE(X_s)]^2 + [SE(X_t)]^2} \ ,$$

where $SE(X_s)$ and $SE(X_t)$ are the standard errors of X_s and X_t, respectively.

The above approximation applies to estimated totals and should not be used for combining estimates of means or ratios. For the latter types of estimates, the appropriate formula must include terms representing the proportion of the population that is represented by each of the two component estimates. For example, if Y_s and Y_t are the estimated means for the two subgroups to be combined, then the combined estimate, Y_c, is given by the formula

$$Y_c = P_s Y_s + (1 - P_s) Y_t \ ,$$

and the standard error of Y_c can be approximated by

$$SE(Y_c) = \sqrt{[P_s SE(Y_s)]^2 + [(1 - P_s) SE(Y_t)]^2} \ ,$$

where Ps is the proportion of the combined group that is included in the subgroups. It should be noted that both forms of the standard error given above are approximations that may understate the true standard error of the combined estimate.

Confidence intervals and relative standard errors can be calculated for all statistics derived from MCBS data (e.g., totals, percentages, means, ratios, and regression coefficients). The following section provides a brief explanation of the method used to compute the standard errors for MCBS estimates.

Variance Estimation (Using the Replicate Weights)

The standard errors reported in the detailed tables in this sourcebook reflect the complexity of the MCBS sample design. In many statistical packages, the procedures for calculating variances assume that the data were collected in a simple random sample. Procedures of this type are not appropriate for calculating variances for statistics based on a stratified, unequal-probability, multistage sample such as the MCBS. They could produce overestimates or, more likely, underestimates of the true sampling error.

Because the MCBS has a complex design, standard errors in the sourcebook tables were estimated with WesVarPC, a statistical software package that accounts for survey design. Estimates of standard errors from WesVarPC are produced using "replication" methods. The basic idea behind the replication approach is to use variability among selected subsamples, or replicates, to estimate the variance of the "full-sample" statistics. These methods provide estimates of variance and standard errors for complex sample designs that reflect weighting adjustments such as those implemented in the MCBS. Replication techniques can be used where other methods are not easily applied, and they have some advantages even when other methods can be used.

Replicate weights for MCBS data have been computed using Fay's variant of Balanced Repeated Replication (BRR). BRR is generally used with multistage, stratified sample designs in which two PSUs are sampled within each stratum, possibly with unequal probabilities of selection. The replicate samples are half-samples formed by selecting one of the two PSUs from each stratum. For BRR, the weights for units in the selected PSUs in each half-sample are doubled and the weights for units in the nonselected PSUs are set to zero. Each replicate consists of a different half-sample; however, it is not necessary to form all possible half-sample replicates, since the information from

all possible replicates can be captured by using a smaller number of "balanced" half-samples. Fay's method is a variant of BRR, in which the sample weights are adjusted by factors between 0 and 2. With a judicious choice of the perturbation factor, Fay's method provides good estimates of standard errors for a variety of statistics. (For more information on Fay's method, see D. Judkins, 1990, "Fay's Method for Variance Estimation," Journal of Official Statistics, 6: 223-240.)

Replicate weights in the 2007 Cost and Use file are named WEIGHT 1,...,WEIGHT100. These replicate weights can be used in WesVarPC to estimate standard errors for MCBS variables. WesVarPC (Version 2) is available at the Westat website—www.westat.com. Documentation for WesVarPC is also provided there. Alternatively, WesVar Complex Samples, which is an enhanced version of WesVarPC, can be purchased directly from SPSS. Descriptions of both packages are available on the website.

An alternative to WesVar is for the user to write a small custom program using a very simple algorithm. If X_0 is an estimate of a parameter of interest formed using the full-sample weights and $X_1,...,X_{100}$ are estimates (calculated by the user) of the same statistic using the corresponding 100 replicate weights, then the estimated variance of X_0 is

$$Var(X_0) = \frac{2.04}{100} \sum_{i=1}^{100} (X_i - X_0)^2 \ .$$

A third option is to use another software package such as SUDAAN (Professional Software for Survey Data Analysis for Multi-stage Sample Designs) to compute population estimates and the associated variance estimates. Two variables, SUDSTRAT and SUDUNIT, have been included in the 2007 Cost and Use file for users of SUDAAN.

Estimates of Net Change

Estimates of net change from year to year can be obtained simply by computing the difference between two "cross-sectional" estimates, i.e., subtracting the 2006 estimate from the 2007 estimate. Each "cross-sectional" estimate is computed by using weights and sample data from the Cost and Use Data File for a particular year.

Computation of standard error estimates of net change is complicated by the fact that the two samples are not independent. Many sample persons are retained in the MCBS sample from year to year. The sample design for selecting each new supplement also uses the same PSUs and many of the same secondary sampling units (SSUs).

Direct Methods. One method for estimating the variances of the differences, when samples are not independent, involves direct estimation of the variances using WesVarPC or SUDAAN. Records from 2 or more years are concatenated into a single file, which retains every record from each of the original files. The user will need to supply instructions to the application to define a variable that represents the difference. The form of these instructions will depend on the particular application package.

In WesVarPC, the "Function" procedure within "Tables" allows a variable to be defined, e.g., net difference between 2006 and 2007 estimates, *d*0706=*cy*07e – *cy*06e. Standard errors associated with estimates of *d*0706 are the required standard errors of the difference.

In SUDAAN, estimates of year-to-year differences can be generated using the CONTRAST option, where the cells to be contrasted are the estimates for each year. This can be accomplished by adding the following statement to the run request:

```
CONTRAST "original file designator" (1, -1)
```

where "original file designator" is the variable that indicates the file in which the record originated (e.g., CY). Standard errors associated with the contrast are the required standard errors of the differences.

For a custom program, the standard errors can be computed using estimate differences for each replicate using the following formula

$$Var(D_0) = \frac{2.04}{100} \sum_{i=1}^{100} (D_i - D_0)^2,$$

where D_0 is the difference between full-sample estimates for each year, and $D_1,...,D_{100}$ are corresponding differences for each replicate sample.

Approximations. For screening purposes, shortcut approximations provide another method for estimating the variances of the differences between two estimates. Shortcut approximations consist of two thresholds, which are based on empirical examination of year-to-year correlations. (R.C. Bailey, A. Chu, and J. O'Connell, 1997, "Considerations for Analysis of the Medicare Current Beneficiary Survey (MCBS) Across Time," ASA, Proceeding of the Section on Survey Methodology, August, 1997.)

The larger threshold, T_L, indicates the minimum absolute difference that may be considered to be significant (at the 5% level). This value is defined as

$$T_L = 2 \cdot \sqrt{V(e_1) + V(e_2)} \ ,$$

where $V(e_1)$ = Var (cy07e) and $V(e_2)$ = Var (cy06e). All differences larger than this in absolute value are considered to be significant.

The smaller threshold, T_S, indicates the maximum absolute difference that is considered to be not significant (at the 5% level). This value is defined as

$$T_S = 2 \cdot \sqrt{\min\left(V(e_1), V(e_2)\right)} \ .$$

All differences smaller than this in absolute value are considered to be not significant. Any difference whose absolute value is between T_S and T_L is indeterminate. These differences will need to be examined using the procedures for direct estimation.

Additional technical questions concerning WesVar or other aspects of MCBS data and public use files may be directed to:

David Ferraro at Westat, telephone (301) 251-4261

To obtain copies of any of the *1992–2003 Health and Health Care of the Medicare Population*, send requests to:

Yuki Jao at Westat, telephone (301) 610-4801
email yukijao@westat.com

To obtain copies of any of the Access to Care Public Use Files or Cost and Use Public Use Files, send requests to:

Bill Long
Office of Research, Development, and Information, C3-24-07
Centers for Medicare and Medicaid Services
7500 Security Blvd., Baltimore, Maryland 21244-1850
telephone (410) 786-7927