Introduction

The Medicare Current Beneficiary Survey (MCBS) is a continuous, multi-purpose survey of a representative sample of the Medicare population, including both aged and disabled enrollees. The accompanying public use file is the second in a planned series of annual data releases reporting Medicare beneficiaries' use of medical services and the costs associated with that medical care. The Cost and Use File is not limited to MCBS survey data alone. It represents a COMBINATION of survey reported data from the MCBS and Medicare claims and other data from the Health Care Financing Administration's administrative files. However, unlike the previously released 1991 through 1994 MCBS Access User Files which also combine survey reports with bill data, the Cost and Use File has undergone a careful RECONCILIATION process to separately identify health care services reported from both sources, from the bill alone, and from the survey alone. This process has produced a file with a more complete and accurate picture of health services received, amounts paid, and sources of payment. The MCBS is sponsored by the Centers for Medicare and Medicaid Services (CMS) in the Department of Health and Human Services of the U.S. Government. Field data collection is done by Westat Corporation.

Advantages of Combining Survey and Administrative Data

The Cost and Use File brings together survey information that can only be obtained directly from a beneficiary with reliable information on services used and Medicare payments made from administrative bill files. Survey reported data includes information on use and costs of health care services as well as information on supplementary health insurance, living arrangements, income, health status and physical functioning. The survey also collects information on health services not covered by Medicare, most notably, prescription drugs and long term facility care. Medicare bill data includes use and cost information on inpatient hospitalizations, outpatient hospital care, physician services, home health services, durable medical equipment, skilled nursing home services, hospice services, and other medical services. This combination file can support a much broader range of research and policy analyses on the Medicare population than would be possible using either survey data or administrative bill data alone.

Matching Survey and Administrative Data

Use and costs of Medicare covered services are reported on both the MCBS survey and in the Medicare central office billing system. This overlap in reporting from the two sources was used to verify the accuracy of survey reports of health service use. Survey reports were matched with administrative bill data to adjust for survey under-reporting using more complete administrative bill data, and to fill in and correct survey reported payment amounts with more accurate information from bills submitted to and paid by Medicare. (Note that this could only be done for services covered by Medicare such as inpatient hospital services, outpatient hospital services, physician services, home health services, acute skilled nursing facility services, durable medical equipment, and other covered services covered. For health services not covered by Medicare such as prescriptions drugs and long term facility care, there was no independent source to which survey reports could be matched.)

Under-reporting of medical services is an enduring problem in personal interview surveys. While respondents can usually recall significant events like hospitalizations for several months, they often fail to recall more routine care like physician visits after a few weeks. In general, as the time interval between the interview date and the medical event increases, the probability decreases that the event will be recalled and reported in the interview. The MCBS interviews persons three times a year, and the average interview recall period is about 4 months. (More frequent interviews would reduce the recall problem, but it would greatly increase both survey costs and the reporting burden on sample persons). Given normal rates of memory decay and the frequency with which aged and disabled persons use medical care, it was reasonable to assume that matching survey events to administrative bills would be helpful in identifying medical events that the sample person could not recall during the interviews.

Match Results

This survey under-reporting hypothesis turned out to be correct. When over 192,000 paid events in Medicare files for MCBS original sample persons were matched to survey reported events, only 104,000 matching survey records (54%) were found. Some small part of the unmatched 88,000 Medicare records are undoubtedly represented in the 76,000 survey-reported events that could not be matched under the criteria used. However, the 76,000 unmatched survey events would be expected to include a substantial share of events that are not covered by Medicare, and therefore would not be expected to match a Medicare paid claim. In addition, only 16,000 of the 76,000 unmatched survey-reported events have a Medicare payment amount. This suggests that the survey reports seriously understate the number of Medicare services when compared to CMS billing records.

The under-reporting problem was more serious for event counts than for Medicare payments. The 88,000 unmatched Medicare events (46% of the total file) accounted for 25% of total Medicare expenditures suggesting that, on average, the events forgotten in the survey interview were less expensive than those that were remembered and reported. This is consistent with the hypothesis that survey respondents tend to remember major health events better than minor health treatments.

In addition to correcting for events that were completely missed in survey reports, the match also helped to fill in missing Medicare payment amounts and correct Medicare payment amounts that had been reported incorrectly. Of the 104,000 survey events matched to Medicare bill records, Medicare was reported as a payer on 78% of these events, and a Medicare payment amount was reported on 61% of these events. This means that the match and reconciliations generated corrections that:

1. made Medicare a payer of record on the 22% of cases where this information was originally omitted in the survey reports;

2. made it possible to determine the correct Medicare payment amount in the 39% of survey records where this information was omitted.

Not all services could be cleanly and easily matched from the two sources. The match employed "strength of evidence" criteria and "hierarchical algorithms" in order to identify matches, survey reports only, bill file reports only, and a small number of similar events for which it was not clear whether there was duplicate survey and bill reports or not. The methods and criteria used in the match are discussed in more detail in the EVENT LEVEL MATCHING discussion in Section 5 of this manual. In addition, Technical Appendix A, "Computer Matching of MCBS Data With Medicare Claims", presents a full discussion of methods, criteria, and early results.

File Building

In order to get a complete and accurate file of services used and payments made, all 104,000 MATCHED service records should be added to all UNMATCHED 88,000 Medicare CLAIM ONLY RECORDS. In addition, unmatched survey reports, EXCLUDING THE 16,000 RECORDS WITH A Medicare PAYMENT AMOUNT, should be added to the matched and Medicare claim only records. This file will be the most complete and accurate file possible, and this combination minimizes the risk of double counting unmatched records. For a more detailed discussion, see the Event Level Matching discussion in Section 5 of this manual.

Imputing Missing Payment Data

In constructing this file particular attention was paid to making payment data, both the amount paid and the sources of payment, as accurate and complete as possible. In the interview itself, interviewers used Medicare and private insurance explanation of benefits forms to accurately record charges and payments. As noted above, we used Medicare administrative bills wherever possible to fill in or correct the Medicare amount reported by the respondent on the survey. For payment amounts where Medicare bills could not be used for correction, a complex imputation process was used to fill in the estimated payments.

One guiding principle used in payment imputations was to preserve, insofar as possible, all partial reports from respondents. For example, many respondents knew how much they paid out-of-pocket for prescription drugs, but did not know how much supplementary private insurance or other third party payers (e.g. Medicaid, VA, HMO) may have paid for that prescription. The out-of-pocket amount reported by the respondent was kept as reported throughout the imputation process as an anchor, and the missing amounts were filled in around it. The first step was to impute a "target reimbursement" amount, that is, a total for that service that was reasonable based on similar cases in the file. The next step was to check which payers were possible (e.g. private insurance, Medicaid, VA, etc.) based on the insurance information reported on the questionnaire and the person's eligibility for public programs. Finally, a computer intensive iterative imputation technique, which borrowed from both Gibbs sampling and "hot deck" methods, was then used to fill in missing payment data for likely payers up to the target reimbursement amount. Emphasis was placed on creating imputed numbers that were not anomalous. That is, imputed amounts were created to be consistent both in level of payment and the share distribution across payers with other similar cases in the file. The techniques and methods used in the payment imputation are described in more detail in the MISSING PAYMENTS AND PAYERS discussion in Section 5 of this manual. In addition, Technical Appendix B, "Imputation of Medical Cost and Payment Data", provides a detailed discussion of the procedures and criteria used to impute missing payments for prescription drug data.

Supplementing the Sample

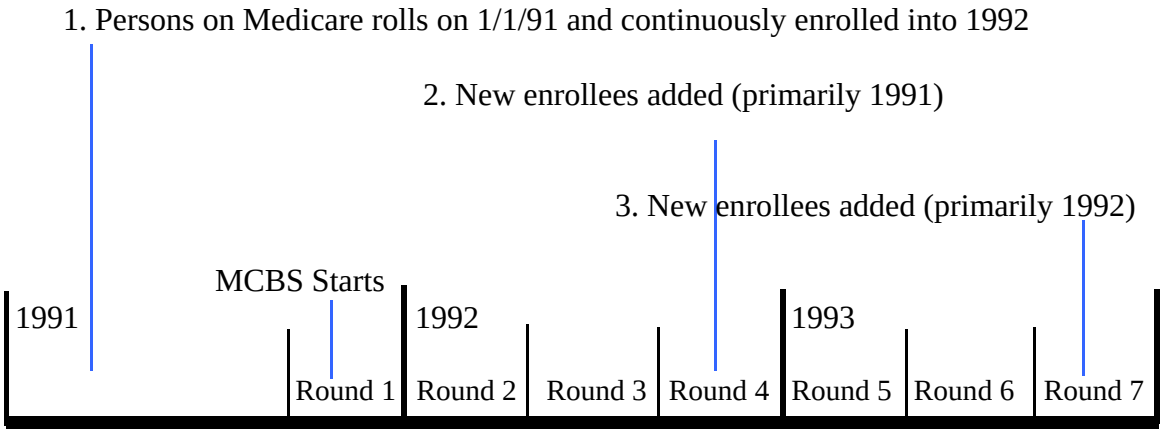
Official Medicare program statistics generally include all persons entitled to Medicare during the year, including those entitled for the entire year, those whose eligibility began during the year, and those who died before the year ended. This mix of continuing enrollees, accretions, and terminations is referred to as "ever enrolled". That is, everyone who was ever enrolled for any time during the year. However, previously released Access To Care User Files from the MCBS represent the "always enrolled", that is, persons continuously enrolled during the entire year. Special steps were needed to

improve the population coverage of the Cost and Use File to the broader concept of "ever enrolled".

The MCBS sample (which is discussed in detail in the SAMPLING AND ESTIMATES section of this report) was drawn from an enrollment list of persons entitled to Medicare on January 1, 1991. This list sample adequately represents persons who were continuously enrolled from January 1, 1991 into 1992. However, it DOES NOT represent persons who became newly eligible for Medicare in 1991 and 1992.

MEDICARE CURRENT BENEFICIARY SURVEY

COMPONENTS OF THE 1992 COST AND USE SAMPLE



NOTES

1. Objective was to add newly enrolled persons not on 1/1/91 sampling list.
2. This creates an 'ever enrolled' population to equate MCBS Cost and Use File estimates with official Medicare program statistics.
3. Since newly added persons were not asked cost and use questions in 1992 (were 'ghosts' added later to the 1992 file), suitable 1992 donors were identified based on their Medicare use profiles to impute their total use and costs.

Each year a supplemental sample is drawn and persons added to the MCBS sample to account for growth in the Medicare population and to replenish the sample for survey persons who died or left the survey during the previous year. This sample replenishment is primarily to insure that each year's MCBS sample adequately represents the entire population.

However, these supplemental samples were also used to add the "missing" newly enrolled persons, that is, those who were on Medicare in 1992 but were not on the January 1, 1991 sampling list. The supplemental sample for Round 4 (September - December, 1992) added persons to the sample primarily to represent those newly enrolled in 1991. The supplemental sample for Round 7 (September - December 1993) added persons to the sample primarily to represent those newly enrolled in 1992. Thus the full sample for the 1992 Cost and Use file is a composite of three groups, those continuously enrolled from January 1, 1991, those newly enrolled in 1991, and those newly enrolled in 1992. The number of persons in each of the three groups, and their collective response rates, are shown in Table 1 in the MCBS SURVEY OPERATIONS section below.

Colloquially, the two groups of newly enrolled are internally referred to as "ghosts", because they were missing, should have been present, but were retroactively added later. Utilization for these persons is included in 1992 data even though they were not actually included in the field sample until late 1992 (Round 4) if they enrolled in 1991 and 1993 (Round 7) if they enrolled in 1992. While interview reports of services used and costs were not available for these "ghosts" in 1992, we did have complete profiles of Medicare use from administrative bill files. To get estimates of total use of services and costs, we matched these "ghosts" to the 1992 file to find appropriate donors based on their Medicare utilization profiles. Once donors were located, the donors use of total (Medicare and non-covered) services and costs were used to impute services and costs for the newly enrolled persons added to the sample in 1991 and 1992. This process brought the sample estimates for persons, use, and costs up to the more complete "ever enrolled" population for Medicare in 1992. A more detailed discussion is included in the SUPPLEMENTING THE SAMPLE discussion of Section 5 of this manual.

Should I Use the Access File or the Cost and Use File?

The Cost and Use File is a MORE COMPLETE file than the previously released Access Files in two fundamental ways:

First, as described above, it includes a more comprehensive definition of the Medicare population. The Access Files sample statistically represents persons continuously in Medicare during the year, the "always enrolled". The first Access file in 1991 excluded persons dying during the year primarily as a matter of necessity, not analytical preference. Since the survey entered the field in September 1991, it was impossible to get a baseline interview with anyone who died between January 1 and August 30.

The Cost and Use File also represents the continuously enrolled, but in addition, represents persons entering the Medicare program during the year, as well as persons dying during the year. This latter group is particularly important to producing accurate total population estimates of spending because use of medical services is generally higher on average in a person's last year of life. Recent internal tabulations of the MCBS sample showed that persons dying in the year are under 5% of the population, but represent over 15% of total expenditures. On a per capita basis, persons dying during the year have spending levels over four times higher than persons continuously enrolled for the entire year.

The second way that the Cost and Use File is more complete than the previously released Access files relates to the services and dollars included in both files. The Access File includes use of services and spending for Medicare covered services only. It is sometimes not recognized that Medicare covers less than half of the average health expenses of its enrollees. (Waldo, Sonnefeld, McKusick, et. Al. "Health Expenditures by Age Group, 1977 and 1987". Health Care Financing Review 10 (4), Summer 1989). The Cost and Use File, by contrast, includes ALL health care services whether covered by Medicare or not. The two most prominent health care services not covered by Medicare, prescription drugs and long term facility care, are included in the Cost and Use File, but are not in the Access file.

File users whose analyses require the ENTIRE MEDICARE POPULATION and/or ALL HEALTH SERVICES WHETHER COVERED BY MEDICARE OR NOT, should use the Cost and Use File rather than the Access File.

File users whose analyses are well served by the CONTINUOUSLY ENROLLED MEDICARE POPULATION and/or MEDICARE COVERED SERVICES ONLY, should use the MCBS Access Files. This includes persons who do year-to-year or LONGITUDINAL ANALYSIS with the 1991 through the 1994 Access Files. For example, a comparison of changes in health status from year to year would be more appropriate using successive annual Access Files. If, for example, the 1992 Cost and

Use File health status information was compared to 1993 Access file information, the results would be confounded because the covered enrollment bases vary. In this situation, it would be very difficult to sort out what part of any 1992 to 1993 differences found were due to genuine year to year trends, and what part to differences between the 1992 ever enrolled and the 1993 always enrolled populations being measured.

Tri-Level File Structure

As an aid to persons using the file, Cost and Use File data is being provided at three different levels of summarization: at the PERSON level, at the TYPE OF SERVICE level, and at the individual EVENT level. The tri-level structure allows analysts to fit the research problem they are addressing to the available file summary levels, and avoid having to process all the detailed event records in the file. For example, an analysis of differences in total health spending per person between men and women could use the person level summary, and thereby avoid having to process the more numerous event level records. Similarly, an analysis of differences in use of Medicare hospital payments by race could use the type of service summary records, and avoid having to process the more detailed event level records. Event level records would be used for more detailed analyses, for example, average length of long term facility stays or average reimbursements per prescription drug. For a more complete discussion of the TRI-LEVEL FILE STRUCTURE, see the beginning of Section 3 of this manual.

MCBS Survey Operations

Fieldwork on the MCBS is conducted for by Westat, Inc., a survey research firm with offices in Rockville, Maryland. Fieldwork for Round 1 began in September 1991 and was completed in December 1991. Subsequent rounds, involving the re-interviewing of the same sample persons or other appropriate respondents, begin every four months. Interviews are conducted regardless of whether the sample person resides at home or in a long term care facility, using the questionnaire version (discussed later) appropriate to the setting.

Repeated Interviews. The MCBS is a longitudinal panel survey. Sample persons are interviewed three times a year over four years to form a continuous profile of their health care experience. The MCBS is thus capable of tracing changes in coverage and other personal circumstances, and observing processes that occur over time, such as people leaving their homes and taking up residence in long term care facilities, or spending down their assets for medical care until they become eligible for Medicaid.

Sample. The MCBS is a stratified random sample of roughly 12,000 beneficiaries selected to be representative of the entire population of aged and disabled beneficiaries enrolled in Medicare in 1992. Sample persons included in the MCBS were sampled from the Medicare enrollment file 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. In order to insure that the sample would yield enough long-term facility stays to produce reliable estimates, some groups of enrollees more likely to enter long term care facilities were over sampled. This included over samples of disabled persons (those under age 65) and very old persons, aged 80 and over.

The sample was drawn from 107 primary sampling units (PSUs) or major geographic areas chosen to represent the nation, including the District of Columbia and Puerto Rico. The sample is annually supplemented during the September through December interview periods (as it was in Round 4 and Round 7) to account for attrition (deaths, disenrollments, refusals, etc.) and newly enrolled persons.

The Community Interview Sample persons in the community (or appropriate proxy respondents) are interviewed using computer-assisted personal interviewing (CAPI) survey instruments installed on notebook-size portable computers. The CAPI program automatically guides the interviewer through the questions, records the answers, and compares them to edit specifications for allowable codes and relationships to other answers. The CAPI thereby increases the amount of accurate and complete information on the front end and lessens the need for after-the-fact editing and corrections. CAPI guides the interviewer through complex skip patterns and inserts follow-up questions where certain data were missing from the previous round's interview. When the interview is completed, CAPI allows the interviewer to transmit the data by telephone to the home office computer.

These interviews yield a series of data over time for each sample person on utilization of health services, medical care expenditures, health insurance coverage, sources of payment (public and private, including out-of-pocket payments), health status and functioning, and a variety of demographic and behavioral information (such as income, assets, living arrangements, family supports, and quality of life). To increase the accuracy of the data collected, respondents are asked to save Explanation of Benefit forms from Medicare, as well as statements from private health insurers and receipts from providers. To assist in accurate reporting of prescription medicines, respondents are also asked to bring to the interview bottles, tubes and prescription bags provided by the pharmacy.

An effort is made to interview the sampled person directly, but in case this person is unable to answer the questions, he or she is asked to designate a proxy respondent, usually a family member or close acquaintance. On average, about 12 percent of each round's community interviews are done by proxy.

The Facility Interview The MCBS conducts interviews for persons in long-term care facilities using a similar, but shortened 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 currently participating in the survey include nursing homes, retirement homes, domiciliary or personal care facilities, distinct long-term units in a hospital complex, mental health facilities and centers, assisted and foster care homes, and institutions for the mentally retarded and developmentally disabled. A complete discussion of how the FACILITY DATA was collected, edited, and formatted into stay records can be found in Section 4 of this manual.

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

Because of the poor health of the long-term facility resident and the preferences of many facility managers that patients not be disturbed, the survey collected information about institutionalized patients from proxy respondents in the facility. In general, nurses or other primary care givers responded to questions about the person's physical functioning and medical treatment. In general, persons from the billing office responded to questions about charges, payments, and sources of payment. The need for interviewers to flexibly switch back and forth among multiple respondents is the primary reason CAPI techniques could not be smoothly used in the facility setting when the survey began. Consequently, traditional pencil and paper techniques were used to collect data for persons residing in long-term care facilities.

The facility instruments include:

(1) The Facility Screener - This instrument gathers information on the facility to determine the facility type and the characteristics of the facility (e.g. size, ownership, etc.). It is asked during the initial interview;

(2) The Baseline Questionnaire - Gathers information on the health status, insurance coverage, residence history, and demographic items on supplemental sample beneficiaries in a facility setting and new admissions from the continuing sample. Selected information from this questionnaire is updated annually for continuing sample persons using an abbreviated version, The Facility Component Supplement to the Core Questionnaire; and

(3) The Facility Core Questionnaire - Collects information on facility utilization, charge and payment information. This questionnaire is asked in every round but the initial one.

Contents of this Documentation

The rest of this manual contains detailed information about this public use file and specific background information intended to make the data more understandable. The sections included are described below.

Section 1: FILE STRUCTURE - Technical description of the public use file specifications and the structure of the public use file. It also provides a brief description and count of each of the record types in this file.

Section 2: CY 1992 COST AND USE CODEBOOK - Codebook of the file variables. This codebook is organized by record type and contains the question number (for data collected in the survey), and variable name, description and location in the record. Codes or possible values and value labels are also supplied. Frequencies for most variables (those with fewer than 120 distinct values) are also included in the codebook, as are notes concerning when variables are inapplicable (that is, questions were not asked due to skip patterns in the CAPI program). An index of variables is also included at the end of the codebook.

Variables in the CMS bill records are documented slightly differently. Record layouts are provided and are cross-walked to CMS data dictionary names. The data dictionary supplies a full explanation of all the variables and their various values.

Section 3: NOTES ON USING THE DATA - Begins with a description of the tri-level file structure, and goes on to describe conventions used to create each separate record (RIC) in the file. This includes notes on how individual variables were collected in cases where variable definitions are not straightforward.

Section 4: EDITS - A list of anomalies that exist in the survey data which were intentionally left as reported by the respondent ("No-Fix" edits), and a description of problems discovered with the CMS administrative data together with the steps taken to correct them. This section also includes a discussion of the creation and editing of Long Term Care Facility stay records.

Section 5: FILLING IN THE GAPS - A detailed description of the adjustments applied to the data to compensate for "missing" information. This includes supplementing the sample list to account for new persons joining Medicare and persons dying during the year, matching survey and administrative data to correct for under-reporting and missing data, and imputation methods to correct for missing payment data and missing payers. Also included is a discussion of the creation and editing of Prescription Drug event records.

Section 6: SAMPLE DESIGN AND ESTIMATION - A description of the MCBS sample design, estimation procedures and projections. A brief discussion of response rates is also included. This section concludes with a comparison of the MCBS projections to CMS control figures.

Section 7: QUESTIONNAIRES - Hard copy versions of the questionnaires used in Round 4. The questionnaires have been annotated with variable names to associate the questions with the codebook. (Even though the data reflect multiple interviews, the Round 4 questionnaires most nearly represent all questions asked in both the introductory and continuing interviews.)

Supplement: CMS National Claims History Data Dictionary, providing information about the claim and bill records.

Technical Appendices: Offer more detail on selected topics in this manual.

A. Imputation of Medical Cost and Payment Data

B. Computer Matching of MCBS Data with Medicare Claims

C. Analytic Edits

D. Setting Source of Payment Flags

Medicare Current Beneficiary Survey
CY 1992 Cost and Use

File Structure

File specifications

The MCBS Calendar Year 1992 Cost and Use public use file consists of a series of separate datasets or files. These datasets contain data on the MCBS sample persons; these files are the **data files.** The other datasets contain SAS® code (SAS input statements, formats and labels) to facilitate the use of the data files by users who have access to a SAS mainframe environment. These are the **README files**.

Figure 1.1a shows file specifications such as file names, record counts, and the associated README file names.

Summary of the Data

The data files represent completed interviews covering calendar year 1992 with a sample of 13,039 Medicare beneficiaries, and supplemental information from CMS's Medicare files. Of these cases, 11,862 beneficiaries were interviewed only in the community, 929 beneficiaries were interviewed only in facilities, and 248 beneficiaries were interviewed in both settings. This release contains full information about the beneficiaries' use of medical services during 1992, and the costs of those services to all payers.

Using the Data

All datasets are standard "flat" files to allow for processing with a wide variety of operating systems and programming languages. The datasets can be divided into three subject matter groups: files related to MCBS survey data with related Medicare administrative variables, files related to cost and use data, and files related to Medicare bill data.

There are several data files containing survey data and related summary administrative variables. For each of these files there is a "README" file which includes a SAS INPUT statement, a PROC FORMAT to interpret the coded fields, LABELs which provide more information about the variable than would be possible in an 8-character name, and a FORMAT statement which associates the code interpretations with the appropriate variables.

Figure 1.1a File specifications

File Name	Records	DCB Information
MCBS.README.RICK	64	RECFM=FB,LRECL=80,BLKSIZE=6160
MCBS.README.RICA	461	(same)
MCBS.README.RIC1	169	(same)
MCBS.README.RIC2	495	(same)
MCBS.README.RIC4	513	(same)
MCBS.README.RIC5	59	(same)
MCBS.README.RIC7	306	(same)
MCBS.README.RIC8	146	(same)
MCBS.README.RIC9	123	(same)
MCBS.README.RICX	228	(same)
MCBS.README.RICDUE	190	(same)
MCBS.README.RICFAE	205	(same)
MCBS.README.RICIPE	193	(same)
MCBS.README.RICIUE	183	(same)
MCBS.README.RICMPE	275	(same)
MCBS.README.RICOPE	172	(same)
MCBS.README.RICPME	248	(same)
MCBS.README.RICSS	83	(same)
MCBS.README.RICPS	114	(same)
MCBS.RICK	13,039	RECFM=FB,LRECL=27,BLKSIZE=8991
MCBS.RICA	13,039	RECFM=FB,LRECL=416,BLKSIZE=8736
MCBS.RIC1	13,039	RECFM=FB,LRECL=63,BLKSIZE=7434
MCBS.RIC2	13,039	RECFM=FB,LRECL=260,BLKSIZE=8840
MCBS.RIC4	13,039	RECFM=FB,LRECL=299,BLKSIZE=7475
MCBS.RIC5	12,109	RECFM=FB,LRECL=23,BLKSIZE=7475
MCBS.RIC7	1,236	RECFM=FB,LRECL=116,BLKSIZE=8932
MCBS.RIC8	42,897	RECFM=FB,LRECL=75,BLKSIZE=9000
MCBS.RIC9	13,039	RECFM=FB,LRECL=140,BLKSIZE=7420
MCBS.RICX	13,039	RECFM=FB,LRECL=851,BLKSIZE=8810
MCBS.RICDUE	11,365	RECFM=FB,LRECL=203,BLKSIZE=7511
MCBS.RICFAE	1,246	RECFM=FB,LRECL=290,BLKSIZE=7250
MCBS.RICIPE	4,529	RECFM=FB,LRECL=229,BLKSIZE=7557
MCBS.RICIUE	664	RECFM=FB,LRECL=218,BLKSIZE=7630
MCBS.RICMPE	264,593	RECFM=FB,LRECL=240,BLKSIZE=7440
MCBS.RICOPE	42,323	RECFM=FB,LRECL=212,BLKSIZE=7420
MCBS.RICPME	181,232	RECFM=FB,LRECL=262,BLKSIZE=8908
MCBS.RICSS	117,351	RECFM=FB,LRECL=238,BLKSIZE=7378
MCBS.RICPS	13,039	RECFM=FB,LRECL=353,BLKSIZE=8825
MCBS.Readme.	1,298	RECFM=FB,LRECL=80,BLKSIZE=9040
MCBS.Billrec.INP	4,247	RECFM=VB,LRECL=3504,BLKSIZE=9000
MCBS.Billrec.SNF	602	(same)
MCBS.Billrec.HSP	188	(same)
MCBS.Billrec.HHA	4,134	(same)
MCBS.Billrec.OTP	27,287	(same)
MCBS.Billrec.PHY	195,377	(same)
MCBS.Billrec.DME	62	(same)

There are several data files containing cost and use data. For each of these files there is a "README" file, which includes a SAS INPUT statement, a PROC FORMAT to interpret the coded fields, and LABELs.

There are seven data files containing Medicare bill data. The MCBS.README.BILLREC file contains SAS input statements and labels (but no formats) for all seven bill record files.

As an illustration of the structure of the README files, Figure 1.2 is a copy of the README file for the Household Composition record, RIC5.

Figure 1.2 Text of a typical README file (MCBS.README04.RIC5 illustrated)

```
INPUT
  @1      RIC             $1.
  @2      FILEYR          2.
  @4      BASEID          $8.
 @12      D_HHTOT         2.
 @14      D_HHREL         2.
 @16      D_HHUNRL        2.
 @18      D_HHCOMP        2.
 @20      D_HHLT50        2.
 @22      D_HHGE50        2.;

PROC FORMAT;

VALUE HHCDFMT    . = 'INAPPLICABLE'
              -8 = 'DONT KNOW'
               1 = 'NO ONE'
               2 = 'SPOUSE ONLY'
               3 = 'SPOUSE & OTHERS'
               4 = 'CHILDREN ONLY'
               5 = 'CHILDREN & OTHERS'
               6 = 'OTHER RELATIVES'
               7 = 'NON-RELATIVES ONLY';

VALUE PEOPLE     0 = 'NO ONE'
               1 = 'ONE PERSON'
               2 = 'TWO PEOPLE'
              20 = 'TWENTY PEOPLE';

COMMENT USE THIS TO SET LABELS ON THE FILE;

LABEL  RIC = 'RIC CODE FOR SURVEY ENUMERATION CODE'
              FILEYR   = 'YY REFERENCE YEAR OF RECORD'
              BASEID   = 'UNIQUE IDENTIFICATION NUMBER'
              D_HHTOT  = 'TOTAL NUMBER OF PEOPLE IN HH'
              D_HHREL  = 'NO. IN HH RELATED TO SP (INCLUDING SP)'
              D_HHUNRL = 'TOTAL NO. PEOPLE IN HH UNRELATED TO SP'
              D_HHCOMP = 'HOUSEHOLD COMPOSITION CODE'
              D_HHLT50 = 'NUMBER IN HH UNDER 50 (MAY INCLUDE SP)'
              D_HHGE50 = 'NO. IN HH 50 AND OVER (MAY INCLUDE SP)';

FORMAT D_HHCOMP HHCDFMT.
              D_HHTOT    PEOPLE.
              D_HHREL    PEOPLE.
              D_HHUNRL   PEOPLE.
              D_HHLT50   PEOPLE.
              D_HHGE50   PEOPLE.;
```

Structure of the MCBS public use file(s)

As mentioned above, the data files can be divided into three subject matter groups: files containing survey data with related Medicare administrative variables, files containing cost and use data, and files containing Medicare bill data.

There are 10 types of records in the survey and administrative summary data group:

- Key
- Administrative Identification
- Survey Identification
- Health Status and Functioning
- Health Insurance
- Household Characteristics
- Facility Characteristics
- Interview
- Residence Time Line
- Cross-sectional Weights

The use and cost records provide detailed and summary information about medical goods and services the beneficiary used in calendar year 1992, the costs associated with those services, and the share of those costs borne by all payers.

There are 15 types of records in the cost and use portion of the file. For some types of utilization, records are provided in two levels of aggregation--detail, and summed by type of utilization.

- Inpatient use and costs (detail and summary)
- Outpatient use and costs (detail and summary)
- Drug use and costs (detail and summary)
- Facility use and costs (detail and summary)
- Dental use and costs (detail and summary)
- Medical services and goods, use and costs (detail and summary)
- Home health use and costs (summary only)
- Hospice use and costs (summary only)
- Person summary of all use and costs

The bill records represent services provided during calendar year 1992 and processed by CMS in conjunction with our administrative functions. To facilitate analysis, the Administrative Identification record contains a summary of the utilization that these bills present in detail. There are seven types of Medicare bill records in the detailed utilization portion of the file:

- Inpatient hospital
- Skilled nursing facility
- Hospice
- Home health
- Outpatient
- Physician/supplier
- DME

All MCBS public use records begin with the same three variables: a record identification code (RIC), the version of the RIC (VERSION) and a unique number that identifies the person who was sampled (BASEID). These elements serve to identify the type of record and to provide a link to other types of records. To obtain complete survey information for an individual, an analyst must link together records for that individual from the various data files using the variable BASEID. In the CY 1992 Cost and Use release, none of the sample people has a record on every data file. Figure 1.3 provides an overview of the presence of data records on the various data files for community and facility respondents.

The tables that follow Figure 1.3 describe all of the types of records in this release. Table 1.A describes the survey and administrative records; Table 1.B describes the bill records.

Figure 1.3 Number of Records present on each of the data files for community and facility respondents

Type of Record	Number of Records present if beneficiary was interviewed in.. Community	Facility	Both settings
RIC K - Key record	1	1	1
RIC A - Administrative Identification	1	1	1
RIC 1 - Survey Identification	1	1	1
RIC 2 - Health Status and Functioning	1	1	1
RIC 4 - Health Insurance	1	1	1
RIC 5 - Household composition	1	0	1
RIC 7 - Facility Characteristics	0	1	1
RIC 8 - Interview Description	1	1	1
RIC 9 – Residence Timeline	1	1	1
RIC X - Cross-sectional weights	1	1	1
RIC DUE - Dental Events	1, several, or none per respondent		
RIC FAE - Facility Events	1, several, or none per respondent		
RIC IPE - Inpatient Hospital Events	1, several, or none per respondent		
RIC IUE - Institutional Events	1, several, or none per respondent		
RIC MPE - Medical Provider Events	1, several, or none per respondent		
RIC OPE - Outpatient Hospital Events	1, several, or none per respondent		
RIC IPE - Prescribed Medicine Events	1, several, or none per respondent		
RIC SS - Service Summary	9 per respondent		
RIC PS - Person Summary	1 per respondent		
Hospital bills *	1, several, or none per respondent		
Skilled nursing facility bills *	1, several, or none per respondent		
Hospice bills *	1, several, or none per respondent		
Home health bills *	1, several, or none per respondent		
Outpatient bills *	1, several, or none per respondent		
Physician/supplier bills *	1, several, or none per respondent		
Durable Medical Equipment bills *	1, several, or none per respondent		

* These bills are summarized in the Administrative Identification record (RIC A), but are provided for more detailed analysis. If the sample person used Medicare benefits, there will be one or many bills, of one or many types, depending on what types of services were used. If the sample person used no Medicare benefits of a certain type, there will be no bills of that type. If the sample person used no Medicare benefits at all, there will be no bills. The RIC A summary provides information about how many services of each type will be found in the bill record files.

Table 1.A: Survey and Administrative Records

File: **KEY**

RIC: "K"

Number of Records: 13,039 - 1 for each person who completed an interview

Description: The BASEID key identifies the person interviewed. It is an 8-digit element, consisting of a unique, randomly-assigned 7-digit number concatenated with a single-digit checkdigit.

In addition to the BASEID, the KEY file contains the type of interview conducted and other variables for classifying the beneficiary.

File: **ADMINISTRATIVE IDENTIFICATION**

RIC: "A"

Number of records: 13,039 - 1 for each person who completed an interview

Description: The ADMINISTRATIVE IDENTIFICATION file contains information about the sample person from administrative records maintained by the Health Care Financing Administration. It contains basic demographic information (date of birth, sex), insurance information (Medicare entitlement, Medicaid eligibility, HMO enrollment), and summarizes the sample person's Medicare utilization for 1992.

Table 1.A: Survey and Administrative Records (Continued)

File: SURVEY IDENTIFICATION

RIC: "1"

Number of records: 13,039 - 1 for each person who completed an interview

Description: The SURVEY IDENTIFICATION file contains demographic information collected in the survey. To some extent, it parallels the demographic information provided in the ADMINISTRATIVE IDENTIFICATION file (date of birth and sex, for example). Demographic information that is not available in the CMS records, such as education, income and military service, is also present.

File: HEALTH STATUS AND FUNCTIONING

RIC: "2"

Number of Records: 13,039 - 1 for each person who completed an interview

Description: The HEALTH STATUS AND FUNCTIONING file contains information about the sample person's health, including: self-reported height and weight, a self-assessment of vision and hearing, use of preventive measures such as immunizations and mammograms, avoidable risk factors such as smoking, and a history of medical conditions. Standard measures - activities of daily living (ADLs) and instrumental activities of daily living (IADLs) - also appear in this file.

Table 1.A: Survey and Administrative Records (Continued)

File: HEALTH INSURANCE

RIC: "4"

Number of Records: 13,039 - 1 for each person who completed an interview

Description: The HEALTH INSURANCE file summarizes the health insurance information provided by the sample people. The file provides both annual and monthly indicators of health insurance coverage by Medicare, Medicaid, HMO's, PHI, and other public plans.

File: HOUSEHOLD COMPOSITION

RIC: "5"

Number of Records: 12,109 - 1 for each person who completed a community interview

Description: The HOUSEHOLD COMPOSITION file contains information about the sample person's household. It reflects the size of the household, and the age and relationship of the people in it.

File: FACILITY CHARACTERISTICS

RIC: "7"

Number of Records: 1,236 - 1 for each sample person interviewed in a facility

Description: The FACILITY CHARACTERISTICS file provides general characteristics of the institutions and most of the information from the facility screener. In several cases, more than one sample person resided in the same facility. In these cases the RIC 7 records are redundant (containing all of the same information), and differ only in the BASEID.

Table 1.A: Survey and Administrative Records (Continued)

File: INTERVIEW DESCRIPTION

RIC: "8"

Number of Records: 42,897 - 1 for each interview

Description: The INTERVIEW DESCRIPTION file summarizes the characteristics of the interview, including type of questionnaire, duration, and whether or not the interview was conducted with a proxy respondent.

File: RESIDENCE TIMELINE

RIC: "9"

Number of Records: 13,039 - 1 for each sample person

Description: The RESIDENCE TIMELINE file tracks the movement of individuals between community and facility settings. While the majority of respondents have only one setting throughout the year, the record allows for up to nine occurrences of movement between a community and a facility setting. See Section 3, Notes.

File: SURVEY CROSS-SECTIONAL WEIGHTS

RIC: "X"

Number of Records: 13,039 - 1 for each sample person

Description: The CROSS-SECTIONAL WEIGHTS file provides cross-sectional weights, including general-purpose weights and a series of replicate weights.

Table 1.A: Survey and Administrative Records (Continued)

File: DENTAL EVENTS

RIC: "DUE"

Number of Records: 11,365

Description: Individual dental events for the MCBS population.

File: FACILITY EVENTS

RIC: "FAE"

Number of Records: 1,246

Description: Individual facility events for the MCBS population. There is one record for each stay, which occurred at least partly in CY 1992. Facility events only contain CY 1992 use and cost information.

File: INPATIENT HOSPITAL EVENTS

RIC: "IPE"

Number of Records: 4,529

Description: Individual inpatient hospital events for the MCBS population.

Table 1.A: Survey and Administrative Records (Continued)

File: INSTITUTIONAL EVENTS

RIC: "IUE"

Number of Records: 664

Description: Individual short-term facility (usually SNF) stays for the MCBS population, which were reported during a community interview or created through Medicare claims data.

File: MEDICAL PROVIDER EVENTS

RIC: "MPE"

Number of Records: 264,593

Description: Individual events for a variety of medical services, equipment, and supplies collected in the survey, including: medical provider (MP), separately billing doctor (SD), separately billing lab (SL), and other medical expenses (OM). See Section 3, Notes.

File: OUTPATIENT HOSPITAL EVENTS

RIC: "OPE"

Number of Records: 42,323

Description: Individual outpatient hospital events for the MCBS population.

Table 1.A: Survey and Administrative Records (Continued)

File: PRESCRIBED MEDICINE EVENTS

RIC: "PME"

Number of Records: 181,232

Description: Individual outpatient prescribed medicine events for the MCBS population. See Section 3, Notes.

File: SERVICE SUMMARY

RIC: "SS"

Number of Records: 117,351

Description: Summarization of the seven individual event RICs along with home health and hospice utilization, yielding a total of nine summary records per person. See Section 3, Notes.

File: PERSON SUMMARY

RIC: "PS"

Number of Records: 13,039

Description: Summarization of utilization and expenditures by type of service and summarization of expenditures by payer, yielding one record per person. See Section 3, Notes.

Table 1.B: Bill Records

File: HOSPITAL BILL

RIC: INP

Number of Records: 4,247

Description: Inpatient hospital bills for the MCBS population. These include bills from short stay general hospitals, and long-term hospitals such as psychiatric and TB hospitals. Different provider types are distinguishable. Generally, there is one bill for each stay. Some hospitals, particularly the long-term facilities, may bill on a cyclical basis and several bills may constitute a single hospitalization.

File: SKILLED NURSING FACILITY BILL

RIC: SNF

Number of Records: 602

Description: Skilled nursing facility bills for the MCBS population. These include Christian Science facilities and other skilled nursing facilities. Different provider types are distinguishable. Generally, several bills constitute a period of institutionalization.

File: HOSPICE BILL

RIC: HSP

Number of Records: 188

Description: Hospice bills for the MCBS population. Billing practices vary by provider in that some hospices bill on a cycle (e.g. monthly) so that several bills constitute a period of hospice care; others submit a series of "final" bills.

Table 1.B: Bill Records (Continued)

File: HOME HEALTH BILL

RIC: HHA

Number of Records: 4,134

Description: Home health bills for the MCBS population. Home health agencies generally bill on a cycle, e.g., monthly.

File: OUTPATIENT BILL

RIC: OTP

Number of Records: 27,287

Description: Outpatient hospital bills for the MCBS population. These bills are generally for Part B services that are delivered through the outpatient department of a hospital (traditionally, a Part A provider).

File: PHYSICIAN/SUPPLIER BILL

RIC: PHY

Number of Records: 195,377

Description: Medicare Part B (physician, other practitioners, and suppliers including DME) claims for the MCBS population. These records reflect services such as doctor visits, laboratory tests, X-rays and other types of radiological tests, surgeries, inoculations, certain other services and supplies, and use or purchase of certain medical equipment.

Table 1.B: Bill Records (Continued)

File: DURABLE MEDICAL EQUIPMENT

RIC: DME

Number of Records: 62
Description: Medicare Part B claims for the MCBS population, which involve the use or purchase of certain medical equipment.

Medicare Current Beneficiary Survey
CY 1992 Cost and Use

Codebooks

This section consists of two parts: 1.) a description of the detail records of survey data and summary data from CMS's administrative and claims files, and 2.) a description of bill and claims detail records from CMS's National Claims History (NCH) database. Included in the first part, "Survey and Claims Summary Records", are frequency distributions for all of the variables in the survey data and for the summary CMS data. The second part of this section, "Medicare Claims Records", does not include frequency distributions.

Survey and Claims Summary Records

Using the tables The tables in this section list the variables in each of the records, give their physical location in the record, list their possible values and relate them to the questionnaires or to source CMS files.

The first part of the Medicare Current Beneficiary Survey public use file (that is, the survey and CMS summary data) is made up of several different types of records. The record type (RIC) is shown on the second line both in the middle of the page and on the upper right hand corner for each page within a section. This will enable more rapid access to particular parts of the codebook. The name of the record being described is on the third line in the middle of the page.

> Variable - This column contains the variable names that we have associated with the SAS version of our data files. Since SAS limits variable names to 8 characters, these names are not always immediately meaningful. You can change them to more informative names, but the names in the tables were used to annotate the copies of the questionnaires.

Certain conventions apply to the SAS variable names. All variables that are preceded by the character "D_", such as D_SMPTYP are derived variables. The variables did not come directly from the survey data, but compiled from several survey variables. Variables preceded by the characters "H_" come for CMS source files.

> Col (Column) - This column locates the variable physically in the record.
>
> Len (Length) - This column describes the length of the field of the variable.

Fmt (Format) Name - This column contains two pieces of information about the variable. First, it identifies the format name associated with the variable in the SAS README file for this variable's RIC. Second, it displays the frequency count for possible values of the variable.

Ques # - The column headed "Ques #" contains a reference to the questionnaire for direct variables, or to the source of derived variables. For example, the "Ques #" entry that accompanies the variable ERVISIT in the Access to Care record is "AC1." The first question in the Access to Care portion of the community questionnaire is the one referenced.

This column will be blank for variables that relate to neither the questionnaire nor to CMS source files. These variables, such as the record identification code (variable name is RIC), are usually ones that we created to manage the data and the file.

Table 2.1 lists the abbreviations that may appear in this column when a section of the questionnaire is referenced.

Table 2.1 Abbreviations Used to Identify Sections of the Questionnaire

Community Questionnaire

IN	Introduction
EN	Enumeration
HI	Health Insurance
HS	Health Status and Functioning
DI	Demographics/Income
CL	Closing

Facility Questionnaire (Screener)

FQ	

Facility Baseline Questionnaire

A	Demographics/Income
B	Residence History
C	Health Status and Functioning
D	Health Insurance
L	Tracing and Closing

Ty (Type) - This column identifies the type of variable; that is, numeric (N) or character (C).

Label (Variable label and codes) - In the first line under this column, you will find an explanation of the variable, which describes it more explicitly than would be possible in only 8 letters. These labels are available in README files, if you wish to use them in creating SAS data sets.

All of the possible values of the variable appear in lines beneath that explanation. Associated with each possible value (in the column labeled "Fmt Name") is a count of the number of times that the variable had that value, and, under the column labeled "Label," a short format expanding on the coded value. Formats are also available in the README files.

Certain conventions were used in coding all variables to distinguish between questions that beneficiaries would not, or could not, answer, and questions that were not asked. These conventional codes are: "." or "-1" if the question was not applicable; "-7" if the respondent refused to answer; "-8" if the respondent didn't know the answer; and "-9" if the answer could not be ascertained from the response. With derived variables, a " " (blank) or "." mean that the variable could not be derived because one or more of the component parts was not available.

Many questions were posed to elicit simple "Yes" or "No" answers, or to limit responses to one choice from a list of categories. In these cases, the responses are "Yes" or "No," or one of the codes from the list. In other questions, the respondent was given a list of items to choose from, and all of the responses were recorded. In these cases, each of the responses is coded "Indicated" or "Not indicated."

If a beneficiary responded with an answer that was not on the list of possible choices, it was recorded verbatim. All of the verbatim responses were reviewed and categorized. New codes were added to the original list of options to accommodate narratives that appeared frequently. For this reason, the list of possible values for some variables may not exactly match the questionnaire.

Inapplicable - Each variable is followed by a statement that describes when a question was not asked, resulting in a missing variable. Questions were not asked when the response to a prior question or other information gathered earlier in the interview, would make them inappropriate. For example, if the sample person said he has never smoked (community component, question HS16), he would not be asked if he smokes now (question HS17).

The codebook for the various survey and summary RICs is followed by a Variable Name Index that lists sequentially all variables in the codebook, source of information, pertinent RIC, and page within the codebook.

Medicare Claims Records

Using the tables The tables in the bill detail section describe the Medicare utilization files included on the public use file. There are two sets of tables; they must be considered together in order to interpret the data in this segment.

File Descriptions for Medicare Claims - These record layouts correspond to the seven Medicare utilization files on the public use file(s). The inpatient hospital and SNF bill files are described in the same record layout even though they are in separate datasets.

NCH No. - The number associated with each variable in the public use file bill records and CMS's Data Dictionary (discussed below). The NCH No. can be used to crosswalk from the bill record to the more detailed description in the dictionary.

Variable - The name we have assigned to the data element (variable). Names may be up to eight characters long, and are mnemonic. The variable name links the record layout to the remainder of the bill detail documentation. This name is also the name that we have supplied in the "README" SAS INPUT statement and labels.

Type - The format of the data element, or variable. Singly occurring data fields may be numeric, character or packed-decimal.

Group items may appear more than once, depending on the information that is present in the bill. For example, if several surgical procedures were reported on the bill, each of them would appear as a separate group item. One surgical procedure would translate to a single group item. A counter shows how many of each trailer type are present. For example, the number of ICD-9-CM procedure code groups present on the claim would be indicated by the counter PROCCNT.

Length - The number of bytes physically occupied by the variable in the record.

Format - How the data should be interpreted. For example, date fields may be read as six characters, interpreted as YYMMDD (two-digit year, followed by two-digit month, followed by the two-digit day of the month).

Description - A more complete explanation of what the variable contains. These descriptions can be assigned to variables with the SAS LABEL code that is provided in the "README" file.

Data Dictionary - The CMS National Claims History Data Dictionary is included as a supplement to this documentation. The data dictionary consists of tables, which are maintained by CMS to describe their internal records. They contain standard definitions of the variables in this file and values for all coded variables. Some of the variables referenced in this dictionary do not appear in this file. We have deleted some fields to protect the privacy of those who are participating in the survey.

Medicare Current Beneficiary Survey
CY 1992 Cost and Use

Notes on Using the Data

This section is a collection of information about various data fields present in this public use release. We have not attempted to present information on every survey data field; rather, we concentrated our efforts on data fields where we have something useful to introduce. We start with information, which is relevant across the board (global information). We follow that with specific information on individual data fields, presented in the same sequence as the data fields appear in the codebook.

Tri - Level File Structure

The Cost and Use file has been summarized at three different levels for the convenience of users. Depending on the specific aims of the analysis, it may be possible for users to avoid having to process all the event records (the most detailed record level in the file) to get totals and subtotals. The type of service summary pulls together event records for each person by type of service used. It is designed to aid analysts who are interested in use, costs, and payer distribution of a particular type of service, for example, average Medicare payments for inpatient hospital services per person during the year or a distribution of payers showing the amount spent on prescription drugs during the year.

While these types of analyses can be prepared from the detailed event records, they can be tabulated more easily - processing many fewer records - from the type of service summary prepared for every sample person. The highest level of summarization is total health spending for each person. This level of summarization can be used, for example, for categorical distributions (e.g. $0 - 500, $501 - 1,000, etc.) of Medicare beneficiaries by the amount spent on health care in a year either in total, or by type of service (e.g. outpatient hospital services). Again, these analyses can also be done using the most detailed event level or type of service summaries, but they could be done more quickly and easily by using the summaries fields that have already been provided.

We recommend that one of the first issues a file user addresses (assuming, of course, a clear picture of the analytical objectives and file processing outputs desired) is whether the file has already summarized use, costs, and payment distributions that would serve their analysis.

To restate this in a more structured way, the Cost and Use File Records (RICs) are assembled at three levels:

1. The Event level reports all payers, costs and utilization at the most detailed level available, for example, a doctor visit, a prescription drug, an inpatient hospital stay, a long term facility stay, an outpatient hospital services bill. (For a more complete discussion of the level of detail represented in each type of event record, please see the discussion in the EVENT LEVEL MATCHING description in Section 5 of this manual).

2. The Type of Service Summary level summarizes all payers, costs and utilization for a person at the type of service level. The seven types of service categories shown are described below. In addition, two new records for services without event level records: home health and hospice services, are included in the type of service summaries. Within each type of service record, separate payer totals for 11 different payers are also shown. Payer totals are shown two ways: once summarizing the event level records, and in adjusted form. The adjusted payer totals are needed to account for differences between Medicare covered days in the year and days covered by interview reference periods. (The first total is also adjusted to exclude unmatched survey event records that are considered duplicative of unmatched Medicare bill record events. See MATCHING EVENT LEVEL DATA in Section 5 for a discussion of this issue). For various logistical reasons, these two periods do not align exactly for all persons in the sample. A discussion of this issue and a description of the how use and payments were adjusted to compensate can be found near the end of Section 5 under the heading ADJUSTING FOR MISSING DAYS AND UNDATED SERVICES.

3. The Person Summary level summarizes all payers and costs across service categories and summarizes type of service amounts. These person records show only one total for each type of service and each payer. Again, payment amounts are shown two ways: as tabulated from event records, and adjusted to compensate for Medicare covered days that were not covered by interview reference periods. (The first total is also adjusted to exclude duplicates as discussed above).

Global Information

Missing values Various negative values are used to indicate missing data. For instance, for survey-collected data, a value of -1 indicates that the variable is inapplicable. A variable is generally inapplicable because the question is not appropriate, for example, a question about hysterectomy when the respondent is a male. In this file, the value -1 has been replaced with SAS® standard missing values (blank for character and "." for numeric). Other missing value codes used in the survey (-7 for "refused," -8 for "don't know," and -9 for "not ascertained") were not changed.

Dates Except for dates of birth, which require century indicators, the dates in this public use release have been written as six numeric characters in the following form: YYMMDD (2-digit year, 2-digit month and 2-digit day). Due to the manner in which the responses were given, these dates must be evaluated in parts because one or more of the parts may be missing. For example, a vague response about a particular date (such as, "I know it was in June of last year, but I'm not sure of the exact day") would be coded "9206-8" ("92" for the year, "06" for June, and the code "-8" for "Don't know" for the day).

Narratives Respondents were asked a number of open-ended questions. The respondents answered these questions in their own words, and interviewers recorded the responses verbatim. The interviewer was prohibited from paraphrasing or summarizing the respondents' answers. However, this public use release does not contain narratives. Instead, we have supplied codes that summarize the answer. Often there will be more than one code because the answer included several specific topics.

Key Record (RIC K)

There are 13,039 key records, one for each individual in the file. Each individual has a variable showing whether they had only community days (11,862 respondents), only facility days (929 respondents) or both community and facility days (248 respondents) in 1992.

The facility interview was conducted whenever the sample person was residing in a facility: 1) that contains three or more beds, 2) that is classified by the administrator as providing long-term care, and 3) whose physical structure allows long-term care residents of the facility to be separately identified from those of the institution as a whole. This broad definition allows analysis beyond traditional views of long-term care, that is, nursing home and related care homes having three or more beds and providing either skilled nursing, or rehabilitative or personal care. Analysts can narrow or extend the focus of their studies of facility care by using information from the Survey Facility Identification Record (RIC 7). This record is present for each sample person for whom a facility questionnaire was administered.

TOT_DAYS is the total number of days in 1992 that the respondent was entitled to Medicare. **C_DAYS** is the number of Medicare-entitled days in 1992 that the respondent was living in the community. **F_DAYS** is the number of Medicare-entitled days in 1992 that the respondent was living in a facility.

FIRSTRND is the survey round that the respondent was first interviewed. See the discussion of SUPPLEMENTING THE SAMPLE in Section 5 for a complete discussion of the supplemental sample respondents who entered the survey in rounds 4 and 7.

Administrative Summary Record (RIC A)

Except as noted otherwise, the variables in this record were derived from HCFA's Medicare enrollment database. History records were searched to establish the beneficiary's status for example: age as of July 1, 1992; residence, type of beneficiary, and other status fields are as of December 31, 1991 or their date of death.

Four variables relating to the sample person's age are provided. Date of birth as reported by the respondent during the initial interview is recorded in the RIC 1 - Survey Identification record (**D_DOB**). Date of birth from the Medicare - Social Security Administration records is recorded in the Administrative Identification Record (**H_DOB**). The variable **H_AGE** represents the sample person's age as of July 1, 1992. The variable **D_STRAT** groups the sample persons by **H_AGE**. The variables **H_DOB, H_AGE,** and **D_STRAT** appear in the Administrative Identification record.

In 19923, approximately 4 million enrollees or 11 percent of the Medicare population had their Part B and/or Part A premiums paid by a State agency. This process, called State buy-in, is tracked by CMS and is used as a general proxy for Medicaid participation. The variables that describe this participation (**H_MCSW** and **H_MCDE01** - **H_MCDE12**) were derived through a match with HCFA's enrollment database.

In 1992, approximately 6 percent of the Medicare population receive Medicare benefits through a coordinated care organization (such as an HMO) which contracts directly with CMS to provide those services. Some of the beneficiaries in the MCBS sample belong to such organizations. The variables that describe this membership (**H_GHPSW** and **H_PLTP01 - H_PLTP12)** were derived through a match with HCFA's enrollment database.

Utilization Summary For easier comparison of groups of people by the number and cost of medical services they have received, the Administrative Identification Record also includes a summary of all Medicare bills and claims for calendar year 1992, as received and processed by CMS through February 17, 1995. (See the variables in the Administrative Identification Record from **H_LATDCH** to the end). Individual bill records are supplied as part of this public use release for researchers who wish to look at Medicare bills in detail (i.e., the HOSPITAL BILL, the SNF BILL, the HOSPICE BILL, the HOME HEALTH BILL, the OUTPATIENT BILL, the DURABLE MEDICAL EQUIPMENT BILL and the PHYSICIAN/SUPPLIER BILL).

The utilization summary represents services rendered and reimbursed under fee-for-service in calendar year 1992. If a beneficiary used no Medicare services at all or was a member of a coordinated or managed care plan (such as a risk HMO) that does not submit claims to a fiscal intermediary or carrier, all their program payment summary variables will be empty. If the beneficiary used no services of a particular type (for example, inpatient hospitalization), the variables relating to those benefits will be empty. Empty variables are zero-filled, except as noted in the next paragraphs.

The variable pertaining to the Part B blood deductible, **H_BLDDED**, is always blank. This information is not consistently available from HCFA's present files. An approximation can be derived from the individual bill records.

The variables pertaining to special coverage (lifetime reserve days, **H_RESDAY**, and psychiatric days, **H_PSYDAY**) are always blank. These benefits are applied to the beneficiary once in a lifetime, and they are decremented as they are used. At the current time, CMS files contain a "current balance" of these benefit days rather than a history of their utilization.

Adjustment bills Initial claims submitted by fiscal intermediaries and carriers for services rendered and paid for by Medicare may be modified by later transactions that result in additional submittal of information relevant to payment or utilization for a given event. There are two types of Part A (institutional) adjustment transactions: credit-debit pairs, and cancel-only credit transactions. Both types of transactions cancel out a bill that was processed earlier (the credit bill exactly matches the earlier bill, which can be viewed as an initial debit). The difference between them lies in how (or if) a new debit transaction is applied to show the correct utilization. If the adjustment consists of a credit-debit pair, the new debit is applied immediately because it is submitted as the "debit" half of the pair. If the adjustment is a cancel-only transaction, the debit may be processed at a later date through a separate bill. In some cases, as when the original bill was completely in error, the cancel-only transaction simply serves to "erase" a mistake, and no new debit would be submitted. For this file, the adjustment processing removes the original debit and the credit which cancels it out, leaving only the final, corrected debit.

> [NOTE: A few rare cases of credit bills with no prior debit may be in this file; these records can be dropped from analysis because they are, in effect, canceling out something of which CMS has no record.]

For Part B claims, we summarized only accepted claims (process code is "A"), or adjusted claims if the adjustment concerned money (process code either "R" or "S" and allowed charges greater than $0). If the claim disposition code (DISPCD) was "03" or "63" (indicating a credit), both the credit and the matching debit were deleted.

Utilization summary - Individual fields After adjustments were processed, the bills were summarized following the rules set forth below.

Inpatient hospital bills Utilization is summarized by admissions, days, charges, covered charges, reimbursement amount, coinsurance days, and coinsurance amount. Admissions **(H_INPSTY)** were totaled by sorting the bills in chronological order and counting the first admission in each sequence. Total covered days **(H_INPDAY)** were summed from **COVDAY** in the bill. Total coinsurance days **(H_INPCDY)** were summed from **COINDAY**. Total bill charges and non-covered charges were selected from the revenue center trailer coded "0001"; total charges were summed as **H_INPCHG** and covered charges (total charges less non-covered charges) were summed as **H_INPCCH**. Coinsurance amounts **(H_INPCAM)** were summed from **COINAMTA** in the bill. Reimbursement **(H_INPRMB)** is the sum of **PROVPAY**, organ acquisition costs (if any) and "pass through" amounts. Organ acquisition costs were accumulated from revenue center trailers when the second and third positions of the code were "81". Pass through amounts were calculated by multiplying covered days (**COVDAY** in the bill record) by the pass through per diem (**PTDIEM** in the bill record).

Skilled nursing facility Utilization is summarized by admissions, days, charges, covered charges, reimbursement amount, coinsurance days, and coinsurance amount. Admissions **(H_SNFSTY)** were totaled by sorting the bills in chronological order and counting the first admission in each sequence. Total covered days **(H_SNFDAY)** were summed from **COVDAY** in the bill. Total coinsurance days **(H_SNFCDY)** were summed from **COINDAY**. Total bill charges and non-covered charges were selected from the revenue center trailer coded "0001"; total charges were summed as **H_SNFCHG** and covered charges (total charges less non-covered charges) were summed as **H_SNFCCH**. Total coinsurance amounts **(H_SNFCAM)** were summed from **COINAMTA** in the bill. Total reimbursement **(H_SNFRMB)** is the sum of **PROVPAY**.

Home Health Utilization is summarized by visits, visit charges, and other (that is, nonvisit) charges. If the second and third positions of the revenue center code were 42, 43, 44, 47, 55, 56, 57, or 58, then the units in the trailer (visits) were added to total visits **(H_HHAVST)** and the charges were accumulated as total covered visit charges **(H_HHACCH)**. If the revenue center codes did not indicate visits, the charges were accumulated as other HHA charges **(H_HHACHO)**. Total home health reimbursement **(H_HHARMB)** was summed from the variable **PROVPAY**.

Hospice Utilization is summarized by days, covered charges, and reimbursement amount. Covered hospice days **(H_HSDAYS)** were summed from the bill variable **COVDAY**. Covered charges were selected from the revenue center trailer coded "0001" and summed as **H_HSTCHG**. Total hospice reimbursement **(H_HSREIM)** was summed from the variable **PROVPAY**.

Outpatient Utilization is summarized by bills, covered charges, and reimbursement amount. Total bills were counted as **H_OUTBIL**. Total covered charges were selected from the revenue center trailer coded "0001" and summed as **H_OUTCHG**. Total outpatient reimbursement **(H_OUTRMB)** was summed from the variable **PROVPAY**.

Part B Physician/Supplier claims Utilization is summarized by number of claims, number of line items, submitted and allowed charges, reimbursement, office visits and office visit charges. All claims and individual line items (there can be up to 13 per claim) were counted and summed as **(H_PMTCLM)** and **(H_PMTLIN)**. Submitted charges and allowed charges **(H_PMTTCH)** and **(H_PMTCHG)** were summed from **SUBCRG** and **ALLOWCRG** in the bill. Total reimbursement for Part B claims **(H_PMTRMB)** was summed from the variable **PAYAMT** in the bill.

Office visits and their charges are summed with other services (described above) and as separate categories **(H_PMTVST and H_PMTCHO)**. We summed office visits and office visit charges separately for two reasons. An office visit is a universally understood measure of service use and access to medical care. It also is an accurate measure of levels of service use across separate groups, unlike charge or payment figures which vary depending on the services that have been performed. Office visits are identified by HCPCS codes in the series 90000-90090 and 99201-99215 in the Part B line item trailer group(s).

Survey Identification Record (RIC 1)

"Initial interview" variables Some questions are asked only in the initial interview for an individual and are not asked again during subsequent sessions because the responses are not likely to change. Such questions include "Have you ever served in the armed forces?" and "What is the highest grade of school you ever completed?". Similarly, once the sample person has told us that he or she has a chronic condition (such as diabetes), the interviewer will not ask, "Have you ever been told you have diabetes?" in a subsequent interview. For this reason, the answers to these questions are missing in later rounds for people who have continued in the survey from an earlier round. To maximize the usefulness of this file, we have filled in this missing information from the original Round 1 (or Round 4 or Round 7) interview. Variables that have been reproduced this way are annotated "Initial interview" in this section.

When the complete date of birth was entered (**D_DOB**), the CAPI program automatically calculated the person's age, which was then verified with the respondent. In spite of this validation, the date of birth given by the respondent (**D_DOB**) does not always agree with the Medicare record date of birth (**H_DOB**). In these cases, the sample person was asked again, in the next interview, to provide a date of birth. Some recording errors have been identified this way, but in most cases beneficiaries provided the same date of birth

both times they were asked. In some cases, proxies indicated that no one was exactly sure of the correct date of birth. In general, it is recommended that the variable **H_DOB** be used for analyses, since the CMS date of birth was used to select and stratify the sample. (Initial interview variable)

The VA disability rating (**D_VARATE**) is a percentage and is expressed in multiples of ten; it refers to disabilities that are officially recognized by the government as service-related. (Initial interview variable)

Race categories (**D_RACE**) are recorded as interpreted by the respondent. Categories were not suggested by the interviewer, nor did the interviewer try to explain or define any of the groups. Ethnic groups such as Irish or Cuban were not recorded. (Initial interview variable)

Hispanic (**D_ETHNIC**) includes persons of Mexican, Puerto Rican, Cuban Central or South American or other Spanish culture or origin, regardless of race. Again, these answers are recorded as interpreted by the respondent. (Initial interview variable)

The respondent was allowed to define marital status categories (**SPMARSTA**); there was no requirement for a legal arrangement (for example, separated). (Initial interview variable)

SPCHNLNM: Respondents were asked to report all living children, whether stepchildren, natural or adopted children. (Initial interview variable)

SPHIGRAD: Education does not include education or training received in vocational, trade or business schools outside of the regular school system. This variable only includes years the sample person actually finished. If the sample person had earned a GED, the response was coded “high school--4th year”. If the sample person said he or she earned a college degree in fewer than 4 years, the response was coded “college and graduate school--4 years”. If the sample person attended school in a foreign country, in an ungraded school, under a tutor or under special circumstances, the nearest equivalent or the number of years of attendance was coded. (Initial interview variable)

INCOME: Income represents the best source or estimate of income during 1992. Round 6 represents the most detailed information for 1992 and is used when available. For individuals not completing Round 6, the most recent information available was used. It should be noted that INCOME includes all sources, such as pension, Social Security and retirement benefits, for the sample person and spouse. In some cases the respondent would not, or could not, provide specific information but did say the income was below $25,000 (or, conversely, $25,000 or more).

Health Status and Functioning Record (RIC 2)

The answers in the health status and functioning section of the questionnaire are a reflection of the respondent's opinion, not a professional medical opinion.

Limitations on activities (**FACLMTAC**) and social life (**HELMTACT**) reflect the sample person's experience over the preceding month, even if that experience was atypical.

In the height measurement **HEIGHTIN**, fractions of an inch have been rounded: those one half inch or more were rounded up to the next whole inch, those less than one half inch were rounded down. (Initial interview variable)

In the weight measurement (**WEIGHT**), fractions of a pound have been rounded: those one half pound or more were rounded up to the next whole pound, those less than one half pound were rounded down. (Initial interview variable)

The sample person was asked to recall or estimate, not to measure or weigh himself or herself.

HYSTEREC: "Hysterectomy" includes partial hysterectomies. (Initial interview variable)

Use of other forms of tobacco, such as chewing tobacco, are not relevant to the "smoking" questions (**EVERSMOK** and **SMOKNOW**). Trying a cigarette once or twice was not considered "smoking," but any period of regular smoking, no matter how brief or long ago, was considered smoking. "Now" meant within the current month or so and not necessarily whether the sample person had a cigarette, cigar or pipe tobacco on the day of the interview. Even the use of a very small amount at the present time qualified as a "yes". Stopping temporarily (as for a cold) qualified as a "yes". (**EVERSMOK** is an initial interview variable)

The answers about difficulty with various tasks (**DIFSTOOP, DIFLIFT, DIFREACH, DIFWRITE, DIFWALK**) reflect whether or not the sample person usually had trouble with these tasks, even if a short-term injury made them temporarily difficult.

The questions about various conditions (**OCARTERY, OCHBP, OCMYOCAR, OCCHD, OCOTHART, OCSTROKE, OCCSKIN, OCCANCER, OCCLUNG, OCCOLON, OCCBREST, OCCUTER, OCCOROST, OCCCERVX, OCCBLAD, OCCOVARY, OCCSTOM, OCCKIDNY, OCCBRAIN, OCCTHROA, OCCBACK, OCCHEAD, OCCFONEC, OCCOTHER, OCDIABTS, OCARTHRH, OCARTH, OCAARM, OCAFEET, OCABACK, OCANECK, OCAALOVR, OCAOTHER, OCMENTAL, OCALZHMR, OCPSYCH, OCOSTEOP, OCBRKHIP, OCPARKIN, OCEMPHYS, OCPPARAL** and **OCAMPUTE**) were coded if the sample person had at some time been diagnosed with the condition, even if the condition had been corrected by time or treatment. The condition must have been diagnosed by a physician, and not by the sample person. Misdiagnosed conditions were not included. If the respondent was not sure about the definition of a condition, the interviewer offered no advice or information, but recorded the respondent's answer, verbatim. (Initial interview variables)

IADLs and ADLs "Difficulty" in these questions has a qualified meaning. Only difficulties associated with a health or physical problem were considered. If a sample person only performed an activity with help from another person (including just needing to have the other person present while performing the activity), or did not perform the activity at all, then that person was deemed to have difficulty with the activity.

Help from another person includes a range of helping behaviors. The concept encompasses personal assistance in physically doing the activity, instruction, supervision, and "standby" help.

These questions were asked in the present tense; the difficulty may have been temporary or may be chronic. Vague or ambiguous answers, such as "Sometimes I have difficulty," were coded "yes."

PRBTELE: Using the telephone includes the overall complex behavior of obtaining a phone number, dialing the number, talking and listening, and answering the telephone.

The distinction between light housework (**PRBLHWK**) and heavy housework (**PRBHHWK**) was made clear by examples. Washing dishes, straightening up and light cleaning represent light housework; scrubbing floors and washing windows represent heavy housework. The interviewer was not permitted to interpret the answer in light of the degree of cleanliness of the dwelling.

PRBMEAL: Preparing meals includes the overall complex behavior of cutting up, mixing and cooking food. The amount of food prepared is not relevant, so long as it would be sufficient to sustain a person over time. Reheating food prepared by someone else does not qualify as "preparing meals".

PRBSHOP: Shopping for personal items means going to the store, selecting the items and getting them home. Having someone accompany the sample person would qualify as help from another person.

PRBBILS: Managing money refers to the overall complex process of paying bills, handling simple cash transactions, and generally keeping track of money coming in and money going out. It does not include managing investments, preparing tax forms, or handling other financial activities for which members of the general population often seek professional advice.

HPPDBATH: Those who have difficulty bathing or showering without help met at least one of the following criteria:

- someone else washes at least one part of the body;
- someone else helps the person get in or out of the tub or shower, or helps get water for a sponge bath;
- someone else gives verbal instruction, supervision, or stand-by help;
- the person uses special equipment such as hand rails or a seat in the shower stall;
- the person never bathes at all (a highly unlikely possibility); or,
- the person receives no help, uses no special equipment or aids, but acknowledges having difficulty.

HPPDDRES: Dressing is the overall complex behavior of getting clothes from closets and drawers and then putting the clothes on. Tying shoelaces is <u>not</u> considered part of dressing, but putting on socks or hose is. Special dressing equipment includes items such as button hooks, zipper pulls, long-handled shoe horns, tools for reaching, and any clothing made especially for accommodating a person's limitations in dressing, such as Velcro fasteners or snaps.

HPPDEAT: A person eats without help if he or she can get food from the plate into the mouth. A person who does not ingest food by mouth (that is, is fed by tube or intravenously) is not considered to eat at all. Special eating equipment includes such items as a special spoon that guides food into the mouth, a forked knife, a plate guard, or a hand splint.

HPPDCHAR: Getting in and out of chairs includes getting into and out of wheelchairs. If the sample person holds onto walls or furniture for support, he or she is considered to receive "help from special equipment or aids," since the general population does not use

such objects in getting in and out of chairs. Special equipment includes mechanical lift chairs and railings.

HPPDWALK: Walking means using one's legs for locomotion without the help of another person or special equipment or aids such as a cane, walker or crutches. Leaning on another person, having someone stand nearby in case help is needed, and using walls or furniture for support all count as receiving help. Orthopedic shoes and braces are special equipment.

HPPDTOIL: Using the toilet is the overall complex behavior of going to the bathroom for bowel and bladder function, transferring on and off the toilet, cleaning after elimination, and arranging clothes. Elimination itself, and consequently incontinence, are not included in this activity, but were asked as a separate question, discussed next.

LOSTURIN: "More than once a week" was coded if the sample person could not control urination at all. Leaking urine, especially when the person laughs, strains or coughs, does not qualify as incontinence.

Health Insurance Record (RIC 4)

This record type is a summary of the respondent's health insurance coverage during 1992. There are five monthly indicators that summarize the respondent's health insurance coverage. **D_CARE1 - D_CARE12** specifies type of Medicare coverage: Part A, Part B, or both. **D_CAID1 - D_CAID12** indicates Medicaid eligibility and how we know about it: from the survey, from HCFA's administrative files, or both. To help the respondent answer the questions about Medicaid, the interviewers used the name of the Medicaid program in the state where the sample person was living. **D_PHI1 - D_PHI12** specifies whether the respondent has private health insurance and the source of it: employer-sponsored, self-purchased, both, or unknown source. **D_HMO1 - D_HMO12** indicates whether the respondent was a member of an HMO and what type: private HMO, Medicare HMO, or both. **D_OTH1 - D_OTH12** indicates the number of other public health insurance plans that the respondent has (e.g. VA coverage, PACE plan, state-sponsored drug plan).

In addition to the monthly health insurance variables there are five annual health insurance variables which summarize the monthly variables: **D_CARE**, **D_CAID, D_PHI, D_HMO,** and **D_OTH.**

TOT_PREM is an estimate of the total health insurance premiums paid by the respondent for all their secondary health insurance. **TOT_PREM** was imputed if premium data was missing for one or more policies and the beneficiary had some

community exposure and none of their secondary health insurance policies were HMO plans.

No attempt was made to statistically impute missing premium data for persons who have one or more HMO plans. Where possible HCFA's administrative data on the premium amount which specific HMOs are allowed to charge members was used to fill in missing HMO premium data. If the premium data for one or more policies is missing for a person with HMO coverage, **TOT_PREM** will be missing.

TOT_PREM estimates the premium cost for coverage of the sample person only. If a policy covered more than one person, the premium attributable to the policy were divided by the number of persons covered (**D_COVNM**x).

A private health insurance plan is one that covers any part of hospital bills, doctor bills, or surgeon bills. It does not include any of the following:

- Public plans, including Medicare and Medicaid, mentioned elsewhere in the questionnaire.
- Disability insurance which pays only on the basis of the number of days missed from work.
- Veterans' benefits.
- "Income maintenance" insurance or "Extra Cash" policies which pay a fixed amount of money to persons both in and out of the hospital . These plans pay a specified amount of cash for each day or week that a person is hospitalized, and the cash payment is not related in any way to the person's hospital or medical bills.
- Workers' Compensation.
- Any insurance plans which are specifically for contact lenses or glasses only. Any insurance plans or maintenance plans for hearing aids only.
- Army Health Plan and plans with similar names (e.g., CHAMPUS, CHAMPVA, Air Force Health Plan).
- Dread disease plans which are limited to certain illnesses or diseases such as cancer, stroke or heart attacks.
- Policies which cover students only during the hours they are in school, such as accident plans offered in elementary or secondary schools.

- Care received through research programs such as the National Institutes of Health.

Detailed information is given for up to five health insurance plans in **D_TYPPL**n, **D_BEGPL**n, **D_ENDPL**n, **D_PHREL**n, **D_COVNM**n, **D_COVRX**n, **D_COVNH**n, **D_PAYSP**n, **D_ANAMT**n, **DHMOPL**n, **D_MHMO**n, **D_OBTNP**n, and **D_INDUS**n.

In **D_PHREL1 - D_PHREL5** the "Policy Holder or "Main insured person" is the member of the group or union or the employee of the company that provides the insurance plans. It would also be the name on the policy, if the respondent had it available.

In **D_ANAMT1 - D_ANAMT5** a premium amount was recorded even if the sample person did not directly pay the premium (if, for example, a son or daughter paid the premium). Premium amounts have been annualized, even though the sample person may not have held the policy for the full 12 months.

Household Composition Record (RIC 5)

A household is defined as the group of individuals either related or unrelated who live together and share one kitchen facility. This may be one person living alone, a head of household and relatives only, or may include head of household, relatives, boarders and any other non-related individual living in the same dwelling unit.

Household membership includes all persons who currently live at the household or who normally live there but are away temporarily. Unmarried students away at school, family members away receiving medical care, etc., are included. Visitors in the household who will be returning to a different home at the end of the visit are not included.

Generally, if there was any question about the composition of the household, the respondent's perception was accepted.

Because the date of birth or exact relationship of a household member was sometimes unknown (perhaps because a proxy provided the information), the sum of the variables "number related"/"number not related" (**D_HHREL**/**D_HHUNREL**) or "number under 50" /"number 50 or older" (**D_HHLT50**/**D_HHGE50**) may not equal the total number of people in the household (**D_HHTOT**).

Facility Characteristics Record (RIC 7)

The value of variables representing "number of beds" (**FACLTBED** and **FACTOBED**) will be missing when either there were no beds of that type in the facility, or the question was skipped.

Interview Description Record (RIC 8)

Proxy rules Wherever possible, the community interviews were conducted directly with the sample person. In most cases, the sample person was able to respond to the interview unassisted. In a few cases, a friend or relative assisted the sample person with the interview. The variables **PROXY**, **D_PROXRL**, **RRECHELP** and **D_IHLPRL** provide information about who was interviewed, and how those respondents are related to the sample person.

People who were too ill, or who could not complete the community interview for other reasons were asked to designate a proxy, someone very knowledgeable about the sample person's health and living habits. In most cases, the proxy was a close relative such as the spouse, a son or daughter. In a few cases, the proxy was a non-relative like a close friend or caregiver. The variable **PROXY** indicates whether or not a community interview was conducted with a proxy respondent, and the variable **D_PROXRL** indicates the relationship of the proxy to the sample person. (Since all facility interviews are conducted with proxy respondents, this variable is "missing" for facility cases.)

If the sample person appeared confused or disoriented at the time of the interview, and no proxy could be identified, the interviewer was instructed to complete the questionnaire as well as possible. If the interviewer felt that the respondent was not able to supply reasonably accurate data, this perception was recorded in the interviewer remarks questionnaire and appears in this record as the variable **RINFOSAT**.

"Sample person language problem" was given as a reason for the use of a proxy in 311 interview cases. More often, language problems were addressed without the use of a proxy. Interpreters were used in some cases, and bilingual interviewers used Spanish-language versions of the questionnaires when the respondent preferred to be interviewed in Spanish. There are both English and Spanish versions of the CAPI survey instrument; the variable **LANG** indicates which version was used.

Proxy respondents were always used in nursing homes, homes for the mentally retarded, and psychiatric hospitals. Sample persons were interviewed directly in prisons when that was permitted. The need for a proxy when interviewing respondents in other institutions was evaluated on a case-by-case basis.

In long-term care facilities, the proxy respondents were members of the staff at the facility identified by the administrator. Usually, more than one respondent was used; for example, a nurse may have answered the questions about health status and functioning, while someone in the business office handled questions about financial arrangements.

Other variables Several questionnaires are administered in the facility interview: a personal baseline for individuals in the supplemental sample found to reside in a nursing facility and for new admissions to a facility from the continuing sample; the core and supplement questionnaires for the continuing sample. The facility screener was administered in every case. Please see Section 5 for copies of all of the instruments and for a more detailed description of when each is administered.

Two variables are supplied to further characterize the interview: **LENGTH** contains the length of the interview, in minutes, and **RESTART** indicates whether or not the interview was interrupted. Community interviews are sometimes interrupted to accommodate the respondent's schedule or for other reasons. We did not calculate the duration of the community interview if the interview was interrupted. Facility interviews are conducted with several instruments and often involve a number of respondents. Since nearly all of the facility interviews are interrupted and total duration is difficult to capture (and interpret), **LENGTH** and **RESTART** are always missing for facility interviews.

Residence Timeline Record (RIC 9)

The timeline record tracks situations as a person moves between community and facility settings. The majority of respondents only have one situation which is a community setting for the entire year. However, this record will account for up to nine occurrences of movement between a community and facility setting.

D_SIT1 - D_SIT9 is the starting date of the situation period. **D_CODE1 - D_CODE9** describes the situation: community, facility, *deemed* community, or *deemed* facility. *Deemed* is used for cases where there is a gap in the interview coverage period. **D_FACID1 - D_FACID9** is the facility identifier, where applicable.

STATUS is the respondent's status as of December 31, 1992: living; deceased; living with at least one interview gap in 1992; deceased with at least one interview gap in 1992; respondent is part of the supplemental sample that began the survey in round 4 or 7.

TYPE is a summary of the respondent's situation for the entire year: community, facility, or both.

Cross-sectional Weights Record (RIC X)

Cross-sectional weights apply to the entire file of 13,039 people and can be used for making estimates of the population enrolled in Medicare at any time during 1992 (the "ever enrolled" population).

The records contain variables to permit analysis using Westat's proprietary software, WESVAR, WESREG and WESLOG to compute variance estimates using the replicate weights. In addition, to enable SUDAAN (Professional Software for SUrvey DAta ANalysis for Multi-stage Sample Designs) users to compute population estimates and the associated variance estimates, two variables have been included in this record, **SUDSTRAT** and **SUDUNIT**. Please see Section 6 for a further discussion about weights and estimation using this file.

EVENT LEVEL RICs

Global Information
The following variable descriptions apply to all of the non-PM Event level RICs.

The **SOURCE** specifies the origin of the event [1=event only reported in the survey; 2=event only known through Medicare claim; 3=event reported in survey and matched to Medicare claim].

EVNTNUM is a unique event identifier collected in the survey. EVNTNUMs prefixed by "C" are events "created" only through presence of a Medicare claim [SOURCE=2].

The type that the event was originally reported as is in **OREVTYPE**. In most cases this is the same as the final **EVNTTYPE**; however, some event types are reclassified as a result of the claim type that the event matched or during the imputation process. For example, a respondent may report that he had an outpatient event (OREVTYPE=OP) and the matching process determined that this event matched a physician claim. EVNTTYPE would be changed to MP. Furthermore, an unmatched OP event may "borrow" data from this event to impute incomplete data. EVNTTYPE on the unmatched "beggar" event would be changed to MP, the same EVNTTYPE as its donor.

In addition, survey reported event types of ER (emergency room visits) have all been reclassified because there is no categorization of Medicare claims by emergency room. If the survey reported ER event matches a Medicare claim it is reclassified according to the claim's service type, which in most cases was a physician or outpatient hospital claim. If the ER event was not matched to a Medicare claim it was reclassified depending on its donor's event type.

CLAIMID is a unique claim identifier within service type that links matched survey events with the Medicare claim.

EVBEGYY, **EVBEGMM**, and **EVBEGDD**, **EVENDYY**, **EVENDMM,** and **EVENDDD** are dates from the matched claim, if the survey event is matched. Otherwise they are dates as reported from the survey. **EVENDYY**, **EVENDMM**, and **EVENDDD** are applicable only to EVNTTYPEs of IP and IU. Dental, medical provider, and outpatient hospital event types (RICs DUE, MPE, and OPE) are included in this file if the date of service was in 1992. Inpatient hospital and institutional (SNF) events are included if the discharge date for the visit was in 1992. If there was a discrepancy between the survey-reported date of service and the matching Medicare claim's date of service, the Medicare claim's date was used to determine the year of service.

SITCODE describes the respondent's location at the time of the event: Community or Facility. Events without dates for respondents who have been in both a community and facility setting during the year have a SITCODE of Both. Values of D (deemed Community setting) and G (deemed Facility setting) exist if there are gaps in a respondent's interview coverage period.

AMTTOT is the total reimbursement the provider received for the service. It is the sum of the eleven payer types:

AMTCARE	Amount paid by Medicare
AMTCAID	Amount paid by Medicaid
AMTHMOM	Amount paid by a Medicare HMO
AMTHMOP	Amount paid by a private HMO
AMTVA	Amount paid by the Veterans Administration
AMTPRVE	Amount paid by a private health insurance plan that is employer-sponsored
AMTPRVI	Amount paid by a private health insurance plan that is individually purchased
AMTPRVU	Amount paid by a private health insurance plan whose source is unknown
AMTOOP	Amount paid by the respondent out-of-pocket
AMTOTH	Amount paid by other public health plan(s)
AMTDISC	Amount of uncollected liabilities

AMTPRVU is only applicable to respondents in Facilities because there was no distinction made during the collection of the facility data as to the source of their private health insurance plan.

Each of the eleven payer types has corresponding imputation flags. **IMPSxxx** indicates whether the payer source was imputed. **IMPAxxx** indicates whether the payment amount was imputed. **IMPTOT** indicates whether the total reimbursement to the provider [AMTTOT] was imputed.

AMTCOV is the amount of the total reimbursement [AMTTOT] that is for a Medicare covered service. **AMTNCOV** is the amount of the total reimbursement that is for a non-Medicare covered service. This is particularly relevant for doctor visits where some of the services itemized in the claim are covered by Medicare and some of the services are for non-covered routine care.

Dental Events Record (RIC DUE)

DVBRIDGE, DVCLEAN, DVCROWN, DVEXTRAC, DVFILLNG, DVORTHO, DVOTHER, DVRTCNAL, DVXRAYS are dental service indicator flags collected in the survey.

Facility Events Record (RIC FAE)

There is one record for each facility stay for the respondent. If the respondent left the facility for a period greater than 30 days and returned to the facility a separate stay record was created. **REFBEGYY**, **REFBEGMM**, and **REFBEGDD** is the earliest date in 1992 that the respondent was in the facility. **REFENDYY**, **REFENDMM**, and **REFENDDD** is the last date in 1992 that the respondent was in the facility. **ADMISYY**, **ADMISMM**, **ADMISDD** is the respondent's date of admission to the facility. **DISCHYY**, **DISCHMM**, **DISCHDD** is the respondent's date of discharge from the facility. **STAYDAY**S is the number of days in 1992 that the respondent was in the facility.

BEGSTAT and **ENDSTAT** describe the respondent's situation at the beginning and ending of the reference period.

D_FACID is a unique facility identifier that can be linked to the Facility Characteristics Record (RIC 7) to contain facility-specific information.

AMTCARE is the amount paid by Medicare to the facility that is <u>not</u> included in any of the other Event records. For instance, most doctor visits that occurred while the person is in the facility are found in the Medical Provider Events Record (RIC MPE); however, if the facility reported an amount received by Medicare that exceeded the total Medicare amounts on the Event RICs, the Medicare amount reported by the facility that is in excess of the other events' Medicare amounts is reported here.

AMTTOT is the sum of the seven facility payer types AMTCARE, AMTCAID, AMTVA, AMTPRVU, AMTOOP, AMTOTH, AMTLIFE. Note that according to the above explanation of AMTCARE this amount is not duplicated in the other Event records.

TOTCARE is the total amount paid by Medicare while the person was in the facility. It includes all Medicare amounts [AMTCARE] from other Event records that occurred during the person's facility stay. Additionally it includes any amount reported by the facility that is in excess of the other events' Medicare amounts.

TOTALL is the sum of TOTCARE, AMTCAID, AMTVA, AMTPRVU, AMTOOP, AMTOTH, and AMTLIFE. Given the definition of TOTCARE, it is the total amount paid for the person while he was in the facility. Note that some of this amount may be duplicated in other Event records.

Inpatient Hospital Events Record (RIC IPE)

ODIAGCNT, PRINDIAG, ODIAG1, ODIAG2, DRG, PROCCNT, PROC1, PROV, STATUS, UTLZNDAY, COINDAY, LRDAYS are variables from the matched Medicare claim. See the Claims Documentation in "Section 2: Codebooks Medicare Claims Records" for further explanation of these variables.

Institutional Events Record (RIC IUE)

These are short-term facility stays that were reported either during a Community interview or created through Medicare claims data. They are in most cases Skilled Nursing Facility stays.

As in the Inpatient Hospital Record, **ODIAGCNT, PRINDIAG, ODIAG1, ODIAG2, DRG, PROCCNT, PROC1, PROV, STATUS, UTLZNDAY, LRDAYS,** and **STATUS** are variables from the matched Medicare claim.

Medical Provider Events Record (RIC MPE)

This record type is a combination of medical provider events collected in the survey: medical provider [MP], separately billing doctor [SD], separately billing lab [SL], and other medical expenses [OM]. The **EVNTTYPE** variable distinguishes between these event types. The classifications of **EVNTTYPE**s are determined by how the respondent reported the event during the survey. For example, a respondent may report an MP event type and total costs associated with it. This may match a Medicare claim with a line item

cost for the physician visit and a line item cost for a lab service. In this case there would not be a separate lab [SL] event.

When an event matched a Medicare claim an effort was made to preserve some of the cost classifications that the claim's line items explicate through the HCPCS code. These groupings are found in the variables **PAMTMED** (physician costs), **PAMTSURG** (surgical costs), **PAMTLABX** (laboratory and x-ray costs), **PAMTOM** (other medical costs such as DME), and **PAMTPM** (prescribed medicine costs). These costs are total reimbursements and they sum to AMTTOT. Note that these variables will only have data for matched survey events and claim-only events.
PROVSPEC is as reported in the survey and will only be present for survey reported events.

OMETYPE, **ORTHTYPE**, **ALTRTYPE**, and **OTHRTYPE** are data collected in the survey for OM (other medical expenses) event types.

Outpatient Hospital Events Record (RIC OPE)

FROMDT and **THRUDT** are dates from the matched Medicare claim indicating that this event represents a period of outpatient hospital visits. **ODIAGCNT**, **ODIAG1**, **ODIAG2**, and **ODIAG3** are variables from the matched claim.

Prescribed Medicine Record (RIC PME)

Some of the variables in this record are only applicable in certain situations during the interview.

Variables that are only applicable when the form of the medication is a pill or a patch are:

TABNUM	(Number of Tablets/patches in the container)
STRNNUM1	(Strength Number)
STRNNUM2	(Strength Number 2nd compound, only applicable to compound drugs)
STRNUNI1	(Strength Unit)
STRNUNI2	(Strength Unit 2nd compound, only applicable to compound drugs)

The following variables are asked of the SP when the medication's dosage form is not a pill, a patch, or a suppository.

AMTUNIT (Amount Unit)
AMTNUM (Amount Number)

SUPPNUM is inapplicable unless the dosage form is a suppository.

Often we impute characteristics about the drug to assist in assigning pricing data. **IMPDF** (the imputed dosage form) was only imputed when there was no match between what was reported and the possible dosage forms found in First Data Bank, or if the form was missing. We also changed the value of **PMFORM** when **IMPDF** was present. The imputed strength, **IMPSTNG**, and the imputed amount number, **IMAMTNUM**, were imputed using various criteria and contributed to determining a unit price only.

The following variables are unadjusted totals for "non-ghosts." These totals do not account for any gap days (days not covered by interview).

SUMTOT (Amount paid by all payers)
SUMCARE (Amount paid by Medicare)
SUMCAID (Amount paid by Medicaid)
SUMPHMO (Amount paid by private HMO)
SUMMHMO (Amount paid by Medicare HMO)
SUMVA (Amount paid by VA)
SUMPRVE (Amount paid by insurance -- employer sponsored)
SUMPRVI (Amount paid by insurance -- self purchased)
SUMUNK (Unknown Amount)
SUMOOP (Amount paid out of pocket)
SUMDISC (Discounted Amount)
SUMOTH (Amount paid by other)

SUMMARY RICS

Type of Service Summary Record (RIC SS)

This record summarizes the Event RICs by person. For every person there are nine records: one record for each of the seven event type RICs (Dental, Facility, Inpatient hospital, Institutional, Medical provider, Outpatient hospital, and Prescribed medicines), plus two additional records which are not present at the event level: Home Health services and Hospice services. The records are identifiable by the **EVNTTYPE** variable:

DU	Dental
FA	Facility
HH	Home health
HP	Hospice
IP	Inpatient hospital
IU	Institutional
MP	Medical provider
OP	Outpatient hospital
PM	Prescribed medicine

When summarizing from the Event level to the Type of Service level any survey-reported event that specified Medicare as a payer that was not matched to a Medicare claim was excluded from the Type of Service summary. Our analysis showed that either 1) the survey event's monies are bundled with a Medicare claim that already matched another survey event, or 2) the respondent was incorrect in reporting Medicare as a payer.

The total amount and the eleven payer types are summarized from the Event RICs into the variables **SAMTTOT**, **SAMTCARE**, **SAMTCAID**, **SAMTDISC**, **SAMTHMOM**, **SAMTHMOP**, **SAMTOOP**, **SAMTOTH**, **SAMTPRVE**, **SAMTPRVI**, **SAMTPRVU**, and **SAMTVA**. The total number of events are summed to **SEVNTS**.

Additional events and expenditures for non-Medicare covered services were imputed for part-year respondents and ghosts. The imputed monies were added to the above SAMT variables to create total monies spent in the variables **AAMTTOT**, **AAMTCARE**, **AAMTCAID**, **AAMTDISC**, **AAMTHMOM**, **AAMTHMOP**, **AAMTOOP**, **AAMTOTH**, **AAMTPRVE**, **AAMTPRVI**, **AAMTPRVU**, and **AAMTVA.** The total number of events reported and imputed are in **AEVNTS**. Note that for full-year respondents the SAMT variables will be the same as the AAMT variables.

Person Summary Record (RIC PS)

The Type of Service record is summarized by person to construct the Person Summary record. There is one record per person with the SAMT variables summed across service types in **SAMTTOT, SAMTCARE**, **SAMTCAID, SAMTDISC, SAMTHMOM, SAMTHMOP**, **SAMTOOP**, **SAMTOTH**, **SAMTPRVE**, **SAMTPRVI, SAMTPRVU**, **SAMTVA**, and **SEVNTS.** The AAMT variables are summed across service type in **PAMTTOT, PAMTCARE**, **PAMTCAID, PAMTDISC, PAMTHMOM, PAMTHMOP**, **PAMTOOP**, **PAMTOTH**, **PAMTPRVE**, **PAMTPRVI, PAMTPRVU**, **PAMTVA**, and **PEVNTS.**

Service types are also summarized across payers for AAMT variables in **PAMTDU, PAMTFA, PAMTHH, PAMTHP, PAMTIP, PAMTIU, PAMTMP, PAMTOP**, and **PAMTPM.** Adjusted number of events by service type are summed in the variables **DUAEVNTS, FAAEVNTS, HHAEVNTS, HPAEVNTS, IPAEVNTS, IUAEVNTS, MPAEVNTS, OPEVNTS**, and **PMAEVNTS.**

Claims Records (DME, HHA, HSP, INP, OTP, PHY, SNF)

The following rules were used to select bill and claims records for this file.

- Inpatient bills were included if the **discharge or "through" date** fell on or after January 1, 1992 and on or before December 31, 1992.
- Skilled nursing facility bills were included if the **admission or "from" date** fell on or after January 1, 1992 and on or before December 31, 1992.
- Home health agency and outpatient facility bills were included if the **"through" date** fell on or after January 1, 1992 and on or before December 31, 1992.
- Hospice bills were included if the **admission or "from" date** fell on or after January 1, 1992 and on or before December 31, 1992.
- Physician or supplier claims were included if the **latest "service thru" date** fell on or after January 1, 1992 and on or before December 31, 1992.

About 16 percent of the sample people did not use Medicare reimbursed services in a fee-for-service setting in 1992; consequently, there are no bill records for them in this file. For other individuals in the sample, we have captured bills meeting the date criteria, processed and made available by CMS through January 1995.

Edits

The use of Computer Assisted Person Interviewing (CAPI) expands and intensifies the data editing process. Many of the edits for accuracy, completeness, and reasonableness are performed immediately as the interviewer enters the information reported. Problems arising from miscommunications or data entry errors often are detected and corrected immediately. In addition, since the CAPI computer software structures the interview by bringing up the appropriate next question without making the interviewer search for it, the software prevents most "skip pattern" errors.

Survey Data Edits

As survey information is collected, it is put into a database management system built into the CAPI software. During the interview and subsequent post interview review, the data in the database are subjected to two types of edits. First, logical relationship edits are performed between various segments of the database to ensure the integrity of the whole. Second, subject matter edits are performed to ensure the internal consistency of the data.

Logical relationship edits ensure that the database is sound by checking the links between segments. For example, every medical provider record in the provider segment must be linked to at least one sample person. The provider record alone without this linkage is not useful.

Subject matter edits ensure the internal consistency of the data. These edits are of two types: those that result in changes to the database to create internal consistency and those that do not. Some edits identify internal inconsistencies, which cannot be corrected because it is not clear which entry is correct. These situations are discussed below in the section on "no fix" edits.

Administrative Bill Data Edits

02 Adjustments

In the late 1980's, CMS decentralized Medicare bill processing operations and shifted Medicare claim review functions to nine host sites around the country. Under the operating procedures in place during 1992, when the deductible field on a claim was incorrect, the host site adjusted the Medicare payment field on the claim, notified the fiscal intermediary of the adjustment, and forwarded the claim to CMS. The fiscal intermediary was not required to re-submit the corrected claim. Only the Medicare total

payment field for the entire claim was adjusted. This means that the deductible field for adjusted claims is incorrect in CMS's files. This is a significant problem for the MCBS since Medicare claims data are an integral component of our activities to verify survey reported information and fill-in gaps in the payment data.

Fortunately, we were able to correct the Part B Medicare claims data. Since only the Medicare payment field was adjusted by the host processing center, we were able to use the disaggregated event level payment data on the Part B claims to reconstruct the Medicare allowed charge and develop corrected Medicare payment and deductible fields. We arranged the discrete Medicare events by service date and adjusted the payment data to be consistent with Medicare law.

Outpatient reimbursement

For a period of time outpatient hospital billing records for three western states were incorrectly showing zero reimbursements. To correct outpatient payments for these three states, a factor relating average program reimbursement to covered program charges was developed. It was used to impute a logical reimbursement amount for these records.

Inpatient Hospital Cost Pass-throughs

The Prospective Payment System (PPS) for inpatient hospital services under Medicare pays a set amount per case. However, this payment excludes some hospital expenses, particularly capital, that are reimbursed on cost basis. (Costs are "passed through" for payment). In order to get total Medicare program payments (that is, actual DRG payment + prorated share of pass-throughs) for an inpatient hospital stay, some method of calculating pass through costs for that stay is needed. Ideally, the provider's cost report could be used to create an accurate measure of pass through costs (on a per-diem basis) that could be applied to individual claims or stays. However, this process is very labor intensive and there are very long delays in getting final hospital cost reports.

In place of the final pass through amounts, each claim contained an estimated pass through amount. Total pass through costs were computed by multiplying the estimated cost pass through per-diem by the number of covered days of care to arrive at the prorated share of pass-throughs applied to each individual claim or stay.

Analysis of claims experience for several States showed that the estimated pass-throughs produced using this method were obviously too high. For three states where the amounts were clearly too high, a national average cost pass through per diem was substituted for the incorrect reports.

"No Fix" Internal Consistency Edits

These edits serve as a warning that certain data are not consistent and cannot be made consistent with only the data and interviewers' notes for guides. These edits are described in Table 4.1. A list of the interviews that failed each follows the edit description.

Table 4.1

NF07001 The number of children given in response to question IN14: "Including natural, adopted and stepchildren, how many living children do you have?" (Community component, Introduction); is less than the total number of people listed by name in the roster and identified as "son" or "daughter". (5 cases)

00185083	00185102	00188842	00198635	00200507

NF07002 The sample person has indicated that he or she has trouble doing light housework, but has indicated no problem with heavy housework. (Community component, Instrumental Activities of Daily Living) (35 cases)

00003483	00005209	00009512	00014441	00016289
00020294	00022559	00041205	00041835	00054299
00057525	00057791	00058643	00086230	00092481
00096142	00097367	00097743	00131985	00141574
00147562	00148492	00151044	00162682	00170217
00174138	00174978	00181532	00182412	00182434
00183762	00185823	00190520	00199496	00202351

Facility Stay Records Data

While data was collected using Computer Assisted Personal Interviewing (CAPI) techniques for persons residing in the community, traditional pencil and paper techniques had to be used to collect data for persons residing in long-term facilities. The reasons for not using CAPI technology were tied to the ways interviews were conducted in facilities, and the limitations of early versions of the CAPI software.

In facilities, information about the patient is often collected from a variety of proxy respondents and record sources. In general, nurses or other primary care givers responded to questions about the person's physical functioning and medical treatment. Generally also, persons from the billing office responded to questions about charges, payments, and sources of payment. In addition to requiring multiple respondents, data collection was complicated because medical and billing records were often physically

located in different places. Interviewers often had to move from person to person and place to place within the facility to get a complete interview.

However, early versions of the CAPI data collection computer software worked best going straight through sequentially from beginning to end. At the time the survey was being fielded, there were limits on how flexibly the early CAPI software could switch backward and forward to accommodate information collected out of sequence. These are the main reasons that facility information was collected in a pencil and paper survey. Since the data was collected in the traditional way, there was considerable emphasis placed on carefully editing this data for completeness and accuracy after collection. These edits and data validation processes are described in more detail below.

Facilities Included - The MCBS survey used a broad definition of long term facility care in order to get a full picture of the types of institutions providing care received by the Medicare population. The survey includes licensed nursing homes and other long term care facilities such as retirement homes, domiciliary or personal care facilities, mental health or mental retardation facilities, continuing care facilities, assisted living facilities, and rehabilitation facilities. To qualify for the survey, a facility must have three or more long term care beds, and answer affirmatively to at least one of three questions: does this facility (1) provide personal care services to residents; (2) provide continuous supervision of residents; (3) provide any long term care.

Note that while the MCBS sample is representative of the Medicare population, which uses long-term care facilities, it was not designed to be precisely representative of the universe of long term care providers. A broad definition of long-term care facilities was chosen to pull in all types of organizations that provide residential long-term care. However, no attempt was made to create a dual beneficiary / provider sampling frame to make the sample simultaneously representative of both the Medicare population and the universe of long term care providers. This decision was made in part because of the difficulty in obtaining a stable and reliable list of non-nursing home long term care providers from which to sample. However, the primary reason was that the MCBS is a continuous, longitudinal survey of Medicare enrollees, not providers. Our approach also avoids the multiple weights and other complications inherent in synthesizing national estimates from a dual person/provider sampling frame. The primary sampling distortion of our person sample on the distribution of long-term care facilities in the MCBS relates to each facility's chances of being included. Larger facilities had a greater chance of being included than small facilities, because at any one time there are more persons in a larger facility than in a smaller facility.

Stay Records - The basic event record measuring use of facility services is a "stay" in a nursing home or other long-term care facility. Stays are measured in terms of days of residence in that facility. A stay is the period of time between admission and discharge for one person in one facility. A person who is in a single facility for an entire year

represents one stay. A person who also spends the year in facilities, but spends the first six months in one facility and then transfers to another, has two stays. A stay begins when a person enters the facility, even if the admission occurred prior to 1992. A stay ends when a person is discharged from a facility or the calendar year ends. Note: This means that all persons in a facility at the end of calendar year 1992, of necessity, will have their stays truncated prior to discharge.

There are some occasions when a person leaves a long-term care facility but is not considered discharged for purposes of creating a stay record. A person residing in a facility who enters a hospital, stays 30 days or less in the hospital, then returns to resume residence in the same facility, does not break their facility stay. A person who goes home for a weekend visit and then returns to the facility, also does not break their stay. However, if a person is formally discharged back into the community, their stay ends.

Need for a Uniform Definition - A period of long term facility residence interrupted only by a brief period of acute hospital care is more accurately characterized as a continuous single episode of long term facility care treatment, rather than two shorter facility "stays" sandwiched around a hospitalization. Unfortunately, there is limited uniformity across nursing homes and other long-term care facilities, and across respondents, in defining when a discharge occurs. Some facilities or respondents treat every time the patient leaves the facility, even for a single day, as a discharge. Others may hold beds for a patient in the hospital for 30 days or more (while charging a bed holding fee) without ever formally discharging them.

These variations across facilities in their patient discharge rules introduce variations in measurements of admissions, discharges, average length of stay, and average payments per discharge, that have little to do with underlying patterns of long term care use. For example, consider two patients in different nursing homes with identical long-term facility use profiles - 67 days between facility admission and discharge with an embedded hospital stay of 7 days in the middle. Nursing home A holds the bed for the first patient during the hospitalization, and calls it one stay that is 67 days long. Nursing home B, by contrast, considers the patient discharged when they go into the hospital, and newly readmitted when they return. It would call this identical facility use two stays, each 30 days in length. A uniform stay definition has been imposed on the MCBS data in order to remove the effect of idiosyncratic discharge policies from the data. The uniform definition allows internally consistent comparisons of facility use across nursing homes with different discharge policies.

File Editing - The facility file has undergone three distinct levels of editing and consistency checking to insure that the information is as accurate and complete as possible. There was no computer driven statistical imputation process used to fill in missing data. Where possible, missing data was filled in using other information from survey responses. For example, if a year-long stay had facility payment records for the

first 10 months of the year, but payments for November and December were missing, the average monthly payment over the 10 reported months was used to fill in the two missing months. This case-by-case editing approach was used judiciously, and primarily for missing payment data (see 3 below).

1. FIELD BY FIELD ACCURACY AND COMPLETENESS - The most basic level of editing was to insure that all necessary fields were completed with legitimate coded answers. Omissions, inaccuracies, and inconsistencies between codes were identified and corrected by staff at Westat Corporation, the primary contractor for the survey. Editors were able, when necessary, to send questions to field survey staff about missing or questionable information.

2. GAPS AND OVERLAPS - One of the basic difficulties in creating a stay record is that it must be built from smaller building blocks that do not automatically fit together evenly. The survey is conducted about once every four months on average. Nursing homes and other facilities usually keep their billing records on a monthly basis, generally using a full calendar month for start and end dates. The beginnings and ends of the facility's billing periods usually do not correspond exactly with the survey reference periods, or the patient's admission and discharge dates. Building the stay requires laying the survey-collected information, within and across reference periods, on a time line for the person. The object is to identify and eliminate any gaps when a person's status was not accounted for, and overlaps where records show a person to be in two places at the same time. This process is particularly complicated for persons who are both in the community and the facility during a year. This editing and file building was done by Westat, which had the most complete first hand information on the person's status during the year. When necessary, CMS administrative records showing dates of medical service, e.g. dates of hospitalization, were also used to make the stay records as complete and accurate as possible.

3. EDITING CHARGE AND PAYMENT DATA - One of the most difficult tasks in a facility survey is collecting complete and accurate charge and payment data. Traditionally in hospitals, there was usually a clear separation on the bill between accommodation charges (for room and board) and charges for ancillary services (for diagnosis and treatment). Taken together, hospital accommodation and ancillary charges and payments represented virtually all facility charges and payments for the period the patient was in residence. In today's more competitive environment, large payers are able to negotiate discounts from posted charges. Medicare has been able to implement an all inclusive per case payment system (the DRG system under PPS). However, most hospitals still have the ability to itemize what is and is not included in their charges and payments. Unfortunately, this ability to clearly identify what is and is not included in payments does not apply to nursing homes and long term care facilities.

a. DIFFERING SERVICE BUNDLES - Understanding differences in payments per day between facilities requires being sure that the dollar amounts apply to the same bundle of facility services. Alternatively, knowing what services are included or excluded from charges and payments helps to explain differences between facilities in average payments per day. For example, if the payments per day in Facility A are $20 a day higher than in Facility B, one might immediately hypothesize that Facility A had higher costs, or a larger profit margin. However, if Facility A's payments include coverage of drugs and Facility B's are only for room and board, the original hypothesis about relative costs or profits are suspect because the payments cover different service bundles. In practice, service bundles included in a charge and payment amount can vary widely based on differences in nursing home practices and the patient's insurance status. Facility charges and payments can be all-inclusive, that is, accommodation charges and all necessary ancillary services and treatment are included in the one basic rate. Alternatively, a facility's charges and payments may cover only room and board. In this facility services and supplies such as drugs, therapy, help with specific needs such as lifting and turning, etc. would be billed a la carte. Many combinations of service bundles in between all inclusive and only room and board are possible.

To further complicate payment data collection, payments for drugs and ancillary services such as therapy furnished to patients in facilities are not always made directly to the facility itself. These payments are often made directly to providers who contract with the facility to provide this care. The task of a facility survey, therefore, is two-fold: first to get facility payment data that is as complete as possible, and then to establish what is and is not included in those facility payments. Unfortunately, a series of questions asked to try to determine whether the facility or a private contractor performed the services, and whether the charges were included in facility payments or not, did not work as planned. We were not able to unambiguously establish what services supplied were and were not included in the reported charges and payments. Responses to these questions were so equivocal that the responses have not been included in the stay records.

b. CHARGE DATA - The MCBS attempted to collect facility charge data separately for base accommodation charges and ancillary service charges, following the traditional hospital billing model. For a number of reasons, collection of this data was very unsatisfactory. Some payers, such as Medicare, pay an all-inclusive rate covering both accommodation and ancillary charges. Other payers, such as Medicaid, pay flat rates for differing bundles of facility services based upon the level of care needed by the patient, without any relationship to posted facility charges - which apply mainly to private patients. Other facilities did not have a charge schedule because they do not treat privately insured patients. For these facilities their "charges" were whatever the third-party payor such as Medicaid would pay on behalf of the patient plus any payments collected directly from the patient (often in the form of a Social Security, SSI, or other pension check). To summarize, the quality of the collected facility charge data was so poor and unreliable, that it was decided to exclude it from the stay record file.

c. PAYMENT DATA - The approach used to test the accuracy of payments per day focused on examining outliers, that is, stays with average payments per day that seemed either too high or too low to be credible. Initially, all nursing homes and other long-term care facilities were grouped together in this analysis. A joint Westat/CMS work team looked at the top 5% and bottom 5% of average payments per day. (Based on a SAS PROC Univariate of the entire distribution of stays using the 5% and 95% levels). A number of cases were resolved because of obvious data entry errors. For example, an extra digit added or a digit missing from either payments or days of stay that distorted average payments per day.

In order to refine the outlier analysis, six different provider categories were created. The categories pull together facilities offering the same level or type of care. The narrower facility categories were expected to have more homogeneous payment per day distributions, and thus permit more sharply focused identification of possible payment outliers within each type. The new categories included three nursing home classifications: Medicare certified, Medicaid certified, and other non-certified nursing homes. The remaining three facility categories were for mental health facilities, facilities for treatment of the mentally retarded, and all other facilities. The work team then analyzed the stay records for the top and bottom 5% of average payments per day in each of the six facility categories.

In general, under reporting or low average payments per day seemed to be a significantly greater problem than payments that were too high. In examining individual cases, one important cause of questionably low payments seemed to be missing billing periods. That is, monthly billing periods where the person was in the facility but no payment was recorded. It usually occurred in the billing periods closest to the interview date, presumably because of time lags in the facility's billing and payment process which made the information unavailable at time of interview. Missing billing periods payment values were edited in using available data from within the stay, as described above. In addition, the missing billing period problem seemed so systemic, that ALL stay records were edited and judiciously corrected to edit in estimated payments when billing periods were missing from a stay.

This still left a number of facilities with average payments per day that were too low to be believable compared to the distribution of payment rates for other facilities in their class. In some cases these low payments were in facilities that appeared to receive funding from global budgets rather than third party or private payments, e.g., a local government facility. In order to limit the distortions to overall data from facilities with such questionably low payments, a minimum acceptable payment level was established for each of the six facility categories. This minimum average payments per day was substituted in the record for any stay record, which after editing, still showed average

payments per day below the minimum acceptable amount. The minimum acceptable payments per day by facility category are as follows:

Medicare certified nursing homes $55;
Medicaid certified nursing homes $40;
Other nursing homes $25;
Mental health facilities $20;
Facilities for the mentally retarded $20;
Other facilities $20.

d. SOURCE OF PAYMENT DATA - Medicare administrative billing data were used to fill in some of the gaps in facility payment data. For stays where all or part of the stay was paid by Medicare as SNF benefits, the program amount paid in Medicare records was compared to the amount reported paid by Medicare in the survey. In cases where the survey reported amount was lower, the higher administrative record amount was substituted in the stay record.

Avoiding Potential Duplication in Facility Medicare Payments In our match of survey reported utilization to Medicare administrative bills for community records where there was a disagreement in the Medicare payment amount, we treated the Medicare amount in billing records as the more accurate report and used it as the final Medicare payment amount. (See EVENT LEVEL MATCHING in Section 5 of this manual for a discussion of matching operations). Similarly, we regard Medicare payment amounts on billing records for services while the person was in the facility (for example skilled nursing home and medical provider bills) as more accurate than what the facility respondents said Medicare paid in the interview. However, given the tri -level structure of the file (see Section 3 in this manual for a discussion), this created a conflict in terms of what payment amounts to report in the event level facility stay record. On the one hand, the facility stay records should contain total payments for all payers to create a complete, stand alone facility stay record of all payments. On the other hand, under the tri -level file structure (see Section 3), Medicare payments from billing records while the person was in the facility are kept separately at the facility type of service and person level. The difficulty is that if total payments from the stay record are added to payments from the Medicare bills, this creates DUPLICATION or double counting of Medicare payments.

To make the facility stay record easy to use, and at the same time prevent double counting of Medicare payments to facilities, we include payment variable in the facility stay record which compute Medicare payments to facilities in two different ways.

When RIC FAE is used **with the other MCBS Event RICs** use AMTCARE for Medicare facility payments and AMTTOT for total facility payments.

AMTCARE includes Medicare payments reported by the facility that EXCEED the total Medicare payment amounts for events which occurred during the facility stay (SITCODE = “F” or “G”) which are included in the MPE, IUE and DUE RICs. (Adding AMTCARE to Medicare payments reported in all other event RICs DOES NOT create any duplication).

AMTTOT includes total facility payments for ALL PAYERS BUT MEDICARE. AMTCARE is substituted for reported Medicare payments in order to exclude Medicare payment amounts for events which occurred during the facility stay (SITCODE = “F” or “G”) which are included in the MPE, IUE and DUE RICs.

When RIC FAE is used as a **stand-alone file**, i.e. without the other MCBS Event RICs, use TOTCARE for Medicare facility payments and TOTALL for total facility payments.

TOTCARE - is the greater of either the total Medicare payments reported by the facility or the total Medicare payment amounts for events which occurred during the facility stay (SITCODE = “F” or “G”) which are included in the MPE, IUE and DUE RICs.

TOTALL - includes total facility payments for ALL PAYERS INCLUDING MEDICARE. In computing TOTALL, TOTCARE is used as the total Medicare facility payment amount.

TOTCARE and TOTALL are included for the convenience of users who analyze stay records only and do not want to link to other MCBS event RICs to get total facility payments. If these amounts are added to Medicare and total payments in the other Event Records, DUPLICATION of Medicare payments will occur.

NOTE: Facility payment amounts exclude payments for inpatient hospitalizations even if they the inpatient stay was embedded in the facility stay.

Filling the Gaps

The 1992 Cost and Use File is designed to provide person level data for estimating total use of, and total payments for, all health care services, covered and non-covered, received by Medicare beneficiaries during calendar year 1992.

This section describes the adjustments that were made to the MCBS data to create a complete file. The adjustments made are as follows:

SUPPLEMENTING THE SAMPLE - These adjustments were made at the sample level to include groups of people who are in the target population (all those who were enrolled at any time in Medicare in 1992), but were not represented in the original sample.

PERSONAL AND SOCIO-ECONOMIC CHARACTERISTICS - These adjustments were made at the person level to include descriptive data (demographics, living situation, socio-economic factors) that are missing because different parts of the questionnaire are initially asked, and subsequently updated, in different interviews.

EVENT LEVEL MATCHING - These operations identified services paid for by Medicare that were not reported on the survey and corrected Medicare payment data reported inaccurately on the survey. A discussion of match results and instructions for building a complete file and avoiding duplication is included.

MISSING PAYMENTS AND PAYERS - These adjustments compensate for missing payment data when the sample person did not know how much an event cost and/or how the event was paid for (by whom, and how much by each payer).

PRESCRIPTION DRUGS - Describes the particular problems encountered in creating the prescription drug event file and how missing payment data was handled.

ADJUSTMENTS FOR MISSING DAYS AND UNDATED SERVICES - These adjustments compensate for data that are missing because some periods of time were not covered by interviews and because some types of health services use (particularly prescription drugs and other medical equipment) are undated.

Adjustments made to records in the Cost and Use file are constrained in two ways. First, because CMS administrative data are used to fill in much of the missing information, all adjustments to MCBS use, cost and source of payment data are consistent with CMS administrative data. For example, if CMS records indicate that the beneficiary is dually entitled to both Medicare and Medicaid, then Medicaid must be considered a possible source of payment when source of payment is missing, even if the beneficiary did not volunteer that information. Second, adjusted data must be consistent with other information for the same person. For example, the source of payment for individual events must be consistent with the sample person's health insurance information.

Basic Principles Although a variety of methods were used in making the adjustments, adjustments of all types are governed by some basic principles. First, information reported by the survey respondent is retained, even if it is not complete, unless strong evidence suggests that it is not accurate. For example, a beneficiary may report having paid $5--the "total cost"--for a prescription that is listed at $25 in the drug wholesale price index. Although it is very unlikely to be the true total cost of the drug, the $5 payment remains with the event as the out-of-pocket *share* of the total.

When information is not reported during the interview, CMS administrative data are the first choice as a source of supplemental, or in some cases, surrogate information. Medicare enrollment information (from the enrollment database, EDB) and bill and claims information (from the national claims history repository, NCH) are used to provide missing personal characteristics, forgotten medical events, and missing or unknown cost information, before statistical imputation. Although the EDB and the NCH are the chief sources of missing data, other CMS administrative files provide information about special areas such as drug costs and Medicaid expenditures.

Finally, when payment data are missing, a total payment ("target reimbursement") is established for each event before the component costs are estimated and allocated to the individual sources of payment. The individual sources of payment are based upon the beneficiary's insurance coverage, both what is reported, and what is known from CMS administrative files. The total cost of the event is largely based on Medicare reimbursement levels and empirically established relationships between Medicare payments and the payments made by secondary payers such as Medicaid or supplemental private health insurance.

SUPPLEMENTING THE SAMPLE

This section describes the adjustments made to the sample to include groups of people who are in the target population, but who are not represented in the interviewed population. The targeted population is the "ever enrolled", that is, all persons enrolled in Medicare at any time during calendar year 1992. The 1992 interviewed population

includes people who were on the Medicare rolls by January 1, 1991, but does not include persons enrolled after that date--people who came onto Medicare rolls during 1991 or 1992.

> Note: Also excluded from the MCBS sample are residents of foreign countries and U.S. possessions and territories other than Puerto Rico.

Targeted Medicare Population--the "Ever-enrolled" The Medicare population is a dynamic group that is constantly changing. Every month, some 200,000 previously unenrolled individuals become eligible and entitled to benefits and are enrolled in Medicare. Such entitlement depends upon meeting the requirements of either the aged, disabled or end-stage renal disease provisions of the Social Security Act and filing for benefits.

In a like manner, every month there are about 150,000 individuals leaving the rolls due to death, non-payment of premiums, recovery from disability, voluntary disenrollment, and other reasons. Thus the net Medicare population continues to grow by about 600,000 people each year.

Producing estimates of total utilization and expenditures for all services (events), including Medicare covered and non-covered, requires an "ever-enrolled" target population. That is, the sample must represent all individuals enrolled in either one or both parts (A and/or B) of the program at any time during the calendar year.

Survey Operational Considerations The MCBS sample is a "list" sample; that is, the people who are selected for interviewing are chosen from a list of all Medicare enrollees. The list of enrollees is based on the Medicare enrollment database (EDB), a complete register of Medicare enrollees. The EDB contains all historical enrollment records, and, to the extent that documentation and transactions affecting the status of individuals are up-to-date, it is a current "snapshot" of the enrolled population.

In a retrospective analysis of the population, the dynamic nature of Medicare enrollment poses no particular problem. Enrollees can easily be identified, categorized and counted, and their records examined. For example, studies on the use of Medicare covered services during the last months of life would have no problems identifying persons who died during the year (after allowing a few months for death notices to flow in and be recorded).

Sampling a population for interviews that will be conducted in the future, however, presents difficulties. The surveyor does not know with certainty in advance whether, or when, a person will join the ranks of the enrolled, or alternatively be removed from them. A sample is selected from a sampling frame as current to the date of interest as possible and field interviews are started, in the knowledge that the great bulk of the targeted

population will be covered by the survey, but that adjustments must be made later for those who should have been included but could not be.

In order to be able to estimate calendar year 1992 cost and use data, it was necessary to continue interviewing the original sample and to select and interview an MCBS supplemental sample in the fall of 1991 (September - December). This initial visit with the supplemental sample just prior to 1992 allowed us to do the following:

- introduce ourselves to the supplemental respondents and explain the purpose and procedures of the study;

- leave material to help in the collection of use and payment information;

- gather baseline data on health status and functioning, demographics, health insurance, and household composition to help in the analysis of the use and cost information to be collected later; and

- obtain data on beneficiaries' access to care to compare with the baseline data collected prior to the implementation of physician payment reform in January 1992. (Published as 1991 Access to Care).

The sample for the MCBS 1992 Cost and Use file was selected from enrollees who were entitled to Medicare on or before January 1, 1991. Most of the people enrolled on that date survived or continued to be enrolled during some or all of calendar year 1992. While making up the greater portion of those ever-enrolled during 1992, the population interviewed for the MCBS in 1992 does not include beneficiaries who were newly enrolled in 1991 after January 1 who survived into 1992, and all beneficiaries newly enrolled in 1992.

The first group, "late 1991" enrollees (that is, after January 1, 1991) was precluded because of the need for time to prepare the sample for the field staff. The second group, 1992 enrollees, could not have been known with certainty in the fall of 1991.

Work on the selection of the January 1, 1991 sampling list began in March 1991, with the production of a "snapshot" of the EDB of persons enrolled for one or both parts of the program as of January 1. The selection of a March update of the EDB allowed three months after the fact (that is, three months after January 1, 1991) for transaction activity (applications for benefits, receipt of death notices, terminations due to other reasons, etc.) to be proceeded. Thus, for the original sample, the snapshot was taken as of the end of March 1991, based on status as of January 1.

In mid-summer, the file was shipped to the contractor (Westat Corporation) for the selection of people who meet the selection criteria (outlined in Section 6 of this manual)

to be included in the supplemental sample. After the sample selection, the contractor developed field instructions for the interviewers, loaded identifying information from the EDB records into the CAPI computer programs, and attempted to locate the sample persons. All of these activities required sufficient lead-time to ensure that the operations could be successfully completed. The lead-time need for field survey activities made it impossible to use a later beneficiary file update to select the sample.

Compositional sample We envisioned the target population of the MCBS 1992 Cost and Use file as composed of three groups: persons enrolled as of January 1, 1991; persons newly enrolled in 1991 after January 1 who lived until 1992; and persons newly enrolled in 1992. As described in the previous section, beneficiaries enrolled after January 1, 1991 could not be included in the Round 1 ("main") MCBS sample. Therefore, these beneficiaries were not interviewed about their medical care and expenditures in 1992. We considered two options for estimating the costs incurred by these new enrollees:

increasing the weights of individuals who resembled new enrollees,

including in the file representatives from the supplemental samples.

The solution employed in the MCBS design to yield an "ever-enrolled" population for calendar year 1992 is to make use of the data for additional enrollees added to the survey as supplemental samples in 1992 and 1993. The 1992 and 1993 supplements, in addition to replacing individuals lost to the survey through death, refusal, etc., include some people who became newly entitled to Medicare in the last four months of 1991 and 1992.

Newly enrolled persons can be any age. Typically, the new enrollee is a member of the youngest "aged" group, that is, those age 65-69. Because this group is proportionately under-represented and because as a cohort, members are moving out and into the next age stratum without commensurate replenishment from the next younger stratum, it was decided that re-weighting the characteristics and patterns of utilization of the remaining group could distort the patterns of use of medical services by putting heavy weights on relatively few cases. By adding new persons from the supplemental samples we increased the sample size of persons in the 65 - 69 age group.

While we had no survey data on use of health services for persons in the supplemental samples, we did have information on their use of covered services under Medicare. The final step in adding these persons to the 1992 file was to identify donors based on similar profiles of Medicare use. The entire pattern of use for the donors, including covered and non-covered services was then transferred to the new persons. In this way, newly enrolled in 1991 and 1992, and suitable patterns of health cost and use, were incorporated into the 1992 Cost and Use file.

As shown in Table 1, the Round 1 MCBS sample consisted of 14,530 Medicare beneficiaries, of whom 14,397 survived until 1992, and thus, were available to be included in the 1992 Cost and Use sample.

Table 1	Eligible beneficiaries	Respondents	Response Rate
Round 1	14,397	11,099	77 %
Round 4	1,095	960	88 %
Round 7	1,183	980	83 %
All	16,675	13,039	78 %

The Round 4 supplement included 1,095 beneficiaries who were not eligible for the original sample because they enrolled in Medicare after the original sampling list was prepared. This includes 642 new Medicare enrollees for 1991 and 453 new enrollees in 1992. Similarly, the Round 7 supplement included 1,183 eligible beneficiaries for 1992, 26 new enrollees for 1991 and 1,157 new enrollees for 1992. The 1992 Cost and Use file is composed of interviews conducted with 13,039 beneficiaries from all three groups, for an over-all response rate of 78 percent.

PERSONAL AND SOCIO-ECONOMIC CHARACTERISTICS

This section describes the adjustments that were made in order to include descriptive data (demographics, living situation, socio-economic factors) that are missing because persons in the Cost and Use file were not initially interviewed at the same time.

Beneficiaries in the original (Round 1) sample population were interviewed throughout 1992. Beneficiaries in the 1992 (Round 4) supplemental sample received only the introductory MCBS interview in the fall of 1992; and those in the 1993 (Round 7) supplemental sample were not interviewed until the fall of 1993. Thus, beneficiaries in the 1992 Cost and Use file can be classified into four sub-categories, depending on the type of information available about them:

I Those who were first interviewed in Round 1 (September through December 1991);

II Those who were first interviewed in Round 7 (September through December 1992);

III Those who were first interviewed in Round 7 (September through December 1993); and

IV Those who were never interviewed.

In the initial or introductory interview, we collect demographic information such as the beneficiary's age, gender, race, education, and income. We also ask about living arrangements and health insurance policies. We ask all beneficiaries to evaluate their own general health, and we ask about chronic illnesses and some standard measures of physical functioning. If the beneficiary is institutionalized, we gather information about the facility, such as ownership and certification, and types of services offered.

The questions about the beneficiary's health are repeated each year, in the September through December round. The facility screener is also re-administered in the fall. Income is updated in the May-August round for the prior year. Insurance and household composition are updated every round.

The Cost and Use file contains our best available information for calendar year 1992 for each of the four subgroups. In some cases, we were able to use data from other years to approximate 1992, in other cases, the data were left missing, to be completed by other types of editing or imputation. Table 2. below summarizes the types of data presented in the MCBS file, and identifies the source of each type of available data.

Note that the 1992 MCBS Cost and Use file contains the same CMS administrative data for beneficiaries in all four subgroups. In every case, the file reflects services rendered during the calendar year 1992, as reported on bills received by CMS through February 1995. Other administrative data (reported in the RIC A) include demographics such as date of birth, sex and race; Medicare entitlement dates for 1992; State buy-in (proxy for Medicaid); whether or not the person belonged to a Medicare-contract HMO; and whether or not the person was receiving hospice benefits.

Table 2: Sources of Information for data presented in the 1992 Cost and Use file

Type of Data	Record ID	Group I	II	III	IV
Demographics	RIC 1	1991	1992	1993	Missing
Income	RIC 1	1992	1992	1993	Missing
Health Status	RIC 2	1992	1992	1993	Missing
Insurance	RIC 4	1992	1992	1993	Missing
Facility characteristics (Facility, only)	RIC 7	1992	1992	1993	Missing
Household composition (Community, only)	RIC 5	1992	1992	1993	Missing
Use (events) and costs		1992	Missing	Missing	Missing
HCFA beneficiary data	RIC A	1992	1992	1992	1992
HCFA billing data		1992	1992	1992	1992

The beneficiaries in Group I represent most Medicare beneficiaries, and are the largest group in the 1992 Cost and Use file. Nearly all of the survey data for this group were collected or updated in 1992. Demographic characteristics other than income are an exception because that information was collected in their initial interview, and not updated. This is also the group from whom we collected (in Rounds 2 through 5 of the survey) complete information about their use of medical services in 1992 and the cost of those services.

The beneficiaries in Groups II and III were added to the survey in supplemental samples. Because most of the descriptive data collected in the survey are collected in the initial MCBS interview, the data for Group II (first interviewed in the fall of 1992) are contemporary with those of Group I--they represent 1992. Data for Group III (first interviewed in the fall of 1993) describe these beneficiaries in 1993; while some individual characteristics might change, we reasoned that the beneficiary's own description (even as of a year later) was more likely to be accurate than one derived by strict statistical imputation. Again, income was an exception; it was self-reported, then edited by imputation for 1992 (Group II) or self-reported for 1993 (Group III).

As indicated in the table, we have no survey data about use and cost of medical services for these groups. We do, however, capture extensive data from Medicare claims and bills, which were used to select appropriate donors to impute the missing data.

The beneficiaries in Group IV are people for whom we have no survey data at all. These beneficiaries died before they could be interviewed, but were nevertheless entitled to benefits for some part of 1992. For the 58 people in this group, we selected individuals similar to them in age group, gender and insurance structure, to act as donors. All survey information for these individuals came from the donors; CMS data (that in the RIC A and the bill records) reflect their own experience.

EVENT LEVEL MATCHING

There are two primary objectives in matching survey reports to Medicare administrative bill records: to correct for under reporting of events on the survey, and to correct errors in payment information collected in the survey.

The first step in matching survey reported medical events to Medicare bill records is gathering all events for a person together. Because the MCBS sample is drawn from CMS's Medicare Enrollment Database (EDB), matching the Medicare paid claims and bills with the correct sample person is a reasonably straightforward process. The beneficiary's Medicare number, or health insurance claim number (HICN), is part of the information collected from the EDB when the sample is drawn. The beneficiary's HICN is verified in the first MCBS interview. Prior to the match, Medicare paid claims are retrieved from the Medicare national claims history repository, by HICN. The search file includes all cross-reference numbers and different beneficiary identification codes associated with each beneficiary, ensuring that all bill records are recovered.

Linking and reconciling the retrieved Medicare claims with individual events reported in the survey is a much more complicated process than matching Medicare paid bills with the correct sample person. There is no data element, or combination of elements, that provide a consistent basis for matching survey data to Medicare claims across all types of services. There are significant differences in the ways that medical goods and services are characterized in the survey and in the Medicare claims records.

Neither the MCBS nor CMS claims records capture a consistent set of data elements for all services types. For example, the MCBS does not capture total reimbursement for inpatient hospital services because the respondent is not likely to know that information; it is not typically included on the notice of utilization, and thus, this information cannot be used in matching. In other categories, especially Part B services, the total charge of the service is known because it appears on the explanation of benefits, and it is a key match field. Similarly, CMS claims data do not always have the same data elements for different claims types. The carrier control number for each claim is included in CMS's claims history files, and the MCBS attempts to collect the carrier control number from the sample person's explanation of benefits in the interviews. As a result, this item is

extremely useful in matching survey reported utilization to Part B claims. On the other hand, the intermediary control number (Intermediaries process claims for Part A of Medicare) is not available in CMS's files, so even though it is collected in the survey, this data element is not helpful in matching the survey data to Part A bill records.

Survey-reported utilization In the utilization sections of the MCBS community questionnaire, beneficiaries are asked about all their medical events, including their visits to practitioners of all types, their prescriptions, and any medical equipment or supplies they might use. (Please refer to Section 7 for copies of the survey instruments and exact wording of the questions).

Types of utilization collected in the MCBS

DU Dentist visits, including cleaning, x-rays and repair, purchase or repair of dentures, and orthodontic procedures.

ER Hospital emergency room visits.

IP Inpatient hospital stays.

IU Other short-term institutional stays, such as skilled nursing home stays or rehabilitation hospital stays.

MP Doctor visits, including visits with medical doctors (MD); practitioners such as chiropractors, podiatrists, audiologists and optometrists; mental health professionals such as psychiatrists, psychologists and clinical social workers; therapists such as physical therapists, speech therapists, occupational therapists, and intravenous and respiratory therapists; other medical practitioners such as nurses and paramedics; and other places offering medical care, such as clinics, neighborhood health centers, infirmaries and urgent care centers.

OP Outpatient visits, including visits to the outpatient department or outpatient clinic of a hospital.

HP/HF Home health visits, collected in the survey as visits by professionals or friends. Health professionals include nurses, doctors, social workers, therapists and hospice workers. Friends include persons who do not live with the beneficiary, but help the beneficiary at home with personal care or other daily needs. These persons may be home health aides, homemakers, friends, neighbors or relatives.

OM Other medical expenses, including purchase or rental of a variety of items: eyeglasses or contact lenses and hearing aids; orthopedic items such as canes, walkers, wheelchairs and corrective shoes; diabetic supplies; oxygen supplies and equipment; kidney dialysis equipment; hospital beds, commodes, and disposable supplies such as disposable diapers and bandages.

PM All prescription medications except those provided by the doctor or practitioner as samples and those provided in an inpatient setting.

In addition to these categories, the community survey instrument is also designed to collect some types of utilization that the beneficiary may unintentionally omit. This utilization is captured when the beneficiary's Medicare and private health insurance statements are reviewed, and is classified as SD - separate billing doctor, and SL - separate billing lab. The SD and SL categories typically include such things as anesthesiology administered while the beneficiary was an inpatient, lab tests not done at the doctor's office, and the radiologist's interpretation of an x-ray.

The facility instruments capture similar information about people who are residents of long-term care facilities, including the use of prescribed medicines.

CMS-reported utilization Medicare claims are basically organized by type of provider. The categories of Medicare claims records are as follows:

Inpatient hospital, psychiatric hospital, TB hospital, Christian Science facility and skilled nursing facility bills. Although these records all share the same format, they contain codes that allow them to be separated into these subcategories. For purposes of the match, bills from skilled nursing facilities were separated from the other types of bills, but no further subdivisions were made.

Home health bills.

Hospice bills.

Outpatient hospital bills.

Part B physician/supplier claims for physician services, diagnostic laboratory and radiology, durable medical equipment and some prescription medicines.

Match categories In matching the survey-reported utilization to the Medicare claims data, MCBS staff frequently must match a Medicare claim category to multiple MCBS categories, and vice versa. Although there are some clear relationships between the categories of utilization collected in the survey and CMS claims categories, not all categories match neatly.

Event-level matching is actually a series of matches between different categories of Medicare claims and MCBS service types. In conducting these matches MCBS staff employ different match algorithms, depending on the data elements available for the particular categories being matched. Matches are arranged in sequence, so that the most similar survey-reported and Medicare claims categories are compared first. The following table presents an overview of the categorical matches.

Figure 1. Overview of event category matches conducted during event-level matching

Matches between similar service types
IP to Inpatient hospital bills
MP, OM, SD, SL to Part B physician/supplier
OP to Outpatient hospital bills
IU to SNF bills
DU to Part B physician/supplier claims
ER to Outpatient hospital bills
HF & HP to Home health agency bills

Matches between less similar service types
ER to Inpatient hospital bills
OP to Inpatient hospital bills
IU to Inpatient hospital bills
IP to SNF bills
IP to Outpatient hospital bills
OP to Part B physician/supplier claims
MP, OM, SD, SL to Outpatient hospital bills

Each match algorithm employs a hierarchy of match criteria that are progressively less restrictive. For example, reported doctor visits are initially compared to claims records by physician's name, date of service, and total charge. If there is not an exact match, the algorithm checks for a match on physician's name and date of service, or total charge and date of service. If there is still no match, the program looks for an exact match on physician's name and total charge, with the date of service match relaxed to dates within one week of each other. (Technical Appendix B contains a more complete discussion of the match.)

The match algorithms not only link survey-reported utilization and Medicare claims records, but also code the records to indicate the strength of the link.

MCBS staff designed the match algorithms to allow survey-reported utilization to be linked to multiple Medicare claims, and vice versa, for two reasons. First, multiple links are often valid. For example, a survey-reported doctor visit may be linked to both a

Medicare claim for the physician's services and a Medicare claim for laboratory services connected with the visit. Second, a stronger match may occur later in the series of matches. A survey-reported doctor visit may have a weak link to a Medicare Part B physician/supplier claim and a strong link to a Medicare outpatient claim. MCBS staff use the link-strength indicator to resolve situations where the multiple matches are logically inconsistent.

Hospice bills were excluded from the match because there is no clean "hospice" category in the survey data. Survey-reported prescribed medicine (PM) utilization was excluded from the match because Medicare coverage of drugs is too limited to warrant complicating the match with immense numbers of survey drug records. Facility and home health utilization were matched in only a summary fashion to improve the accuracy of Medicare payment data.

Three outcomes are possible from the attempted match of survey data to Medicare claims data: the information from the two sources agrees (a match); or, information reported in the survey is not present in the Medicare claims data; or, information is present in the Medicare claims data which was not reported in the survey.

Pre-match edits Prior to matching, the Medicare claims data were edited for obvious omissions and inconsistencies. Please see Section 4: Edits, for a description of the edits applied to CMS bill data and to survey data.

1992 Cost and Use file "events" The matching programs produce a set of records, which reflect the best combination of survey and Medicare claims categories, and present records from both sources (matched and un-matched) in a uniform format.

Figure 2. Categories of utilization in 1992 Cost and Use file
Event-level data

PME Prescription medicine (individuals living in the community, only)

IPE Inpatient hospital, including emergency room visits which result in an inpatient admission

OPE Outpatient hospital, including emergency room visits which do not result in an inpatient admission

MPE Medical doctor and practitioner visits, diagnostic laboratory and radiology, medical and surgical services, durable medical equipment and non-durable supplies.

DUE Dental

IUE Institution (other than inpatient hospital, and other than long-term care)

FAE Facility stay records

Person-level data only

Home health
Hospice

Since the categories of utilization in the Medicare claims do not match the survey categories, utilization groups in the 1992 Cost and Use file are a combination of the two sources.

Event level records - The most disaggregate level of utilization records in the 1992 Cost and Use file is the "event" level record. Event records combine survey-reported information and Medicare claims data in the seven categories presented in Figure 2: IPE, OPE, IUE, DUE, MPE, PME and FAE. **Event records contain a variable to indicate the source of the utilization information--Medicare claims data, survey data, or both--and a variable linking the event record to the bill data, if both sources provided the information.**

Event records also provide a consistent analytic unit within a category of utilization. The following definitions apply to events in this file:

Prescription drugs (PME) The basic unit measuring use of prescription drugs is a single purchase of a single drug in a single container.

Inpatient hospital (IPE) The basic unit measuring use of inpatient hospital services is a single admission. If the beneficiary was still hospitalized at the end of the year, the inpatient event record is not complete, but all data through the end of 1992 are present.

Outpatient (OPE) The basic unit measuring use of outpatient services is a separate visit to any part of the outpatient department for a survey-reported event. For Medicare claim only events, it may represent 1) a single visit; 2) multiple procedures or services within one visit; 3) multiple visits billed together.

Medical, surgical and diagnostic services, and equipment and supplies (MPE)
The basic unit measuring use of these services is a separate visit, procedure, service, or a supplied item for a survey reported event. For Medicare claim only events, it may represent 1) single or multiple visits; 2) single or multiple procedures; 3) single or multiple services; 4) single or multiple supplies; depending on the number of items pulled together on the bill.

Dental (DUE) The basic unit measuring use of these services is a single visit to the dentist, at which time a variety of services, including cleaning, x-rays and an exam might be rendered.

Institution (IUE) The basic unit measuring use of these services is an admission. If the beneficiary was still in the institution at the end of the year, the institutional event is not complete, but all data for 1992 are present.

Facility events (FAE) The basic unit record measuring use of facility services is a "stay" in a nursing home or other long term care facility. Stays are measured in terms of days of residence in that facility. If a person is still in the facility at the end of 1992, the stay is not complete, but all data through the end of 1992 are included.

Emergency room The emergency room (ER) survey category was split between IPE and OPE. Under the prospective payment system, emergency room services that result directly in a hospital admission are included in the DRG payment for the inpatient stay, and thus are not associated with any separate charges or claims. Emergency room visits that are not immediately followed by an inpatient admission are classified as outpatient services. For this reason, survey-reported emergency room (ER) utilization was matched to outpatient, then inpatient bill records, and is reflected in the 1993 Cost and Use file as either OPE or IPE records. Several other survey categories (MP, SD, SL and OM) have been combined to make up the single EMP category. Hospice services do not exist as a separate category of utilization in the survey data, so this category derives from the Medicare claims data.

Post-match edits For most types of services, the MCBS collects a date of service to assist in matching survey-reported data to claims records. Respondents may not always recall exact dates, so dates are collected in three independent parts--month, day and year. Since the year portion of a survey date may be missing or incorrect, records for services in 1991 and 1993 were not eliminated from the survey file until the match was concluded. Similarly, respondents may "telescope" events, believing them to have taken place recently when in reality they occurred a year or more in the past. As matching Medicare claims might help to identify and eliminate these responses, the Medicare records were also not edited on date until after the match; for the match records included services rendered in 1991 and 1993, as well as 1992. After matching, the event file was edited to exclude all services that were rendered outside of calendar year 1992.

If the survey-reported data matched Medicare claims data, the dates of service on the Medicare record were carried into the event record. Dates of service were used as a match criterion in most of the matches, so in many cases, the dates of service in the event record did not change from those reported.

SUMMARY OF MATCH RESULTS

A total of 192,666 Medicare bill events for sample persons who were interviewed about their health care use during the time they lived in the community were matched against 179,966 survey reports. A match was recorded for 104,349 event records, which is 54% of total Medicare bill records events and 58% of survey reported events. The percentage of dollars matched was considerably higher. The 88,000 unmatched Medicare bill events represent 46% of events, but only 24% of total payments. That is, 76% of total dollars on the Medicare bill side were successfully matched with survey reports.

Unmatched Medicare events ($259) were less than half as expensive on average as matched events ($671). This is consistent with general household survey experience that major, more expensive medical events, are more likely to be remembered and reported at the interview.

Evidence supporting improved accuracy

On the 104,000 matched events, Medicare should have been reported as a payer on 100% of the survey reported events. However, Medicare was only reported as a payer for 81,100 or 78% of events. Consequently, the match corrected 22% of the records to make Medicare a payer of record.

On the 104,000 matched events, the Medicare payment amount was only reported on 61% of survey reports. The match filled in the correct Medicare payment for the remaining 39% of survey reports.

Examining 63,000 of the 104,000 matched events where both a Medicare payment and total payment was reported:

> the survey reported Medicare payments overstated Medicare payments from Medicare bill records by $5.7 billion;
>
> the survey reported total payments overstated the total payment amounts from Medicare bill records by $16.4 billion;
>
> these erroneous survey reported payment amounts suggest that Medicare paid only 55% of total payments compared to 70% from the Medicare bill record amounts

Evidence of survey under-reporting

The 88,000 unmatched Medicare paid bill events strongly suggest a high level of under-reporting on the survey. While there are 76,000 unmatched survey reports on the other side, many of these events could not be reasonably expected to be undiscovered matches. For example:

Unmatched survey events unlikely to match an unmatched Medicare bill

1. Over 10,000 unmatched survey events were for dental services, which are rarely covered by Medicare.

2. Almost 8,000 unmatched survey events had total payments equal to zero. (These were very likely parts of bundles of services that were covered in one global payment on the Medicare side, for example, post operative services which were covered by a global surgery fee.)

3. Another 5,000 unmatched survey events were for Medicare HMO enrollees. Virtually all of the Medicare services for these persons are paid through a capitated payment amount and the likelihood is very small that their events ever match a fee for service Medicare paid bill record.

4. There were 3,500 unmatched survey events where the sample person was only entitled to Part A or Part B of Medicare, but not both. Therefore a survey reported service could reasonably not be expected to match a Medicare paid bill record.

5. Another 2,200 unmatched events were provided by the Veterans Administration or in a military installation where no Medicare bill would be expected.

6. Over 12,000 unmatched survey events were for other medical services. While Medicare covers durable medical equipment such as wheelchairs and supplies such as oxygen, it does not cover many items in the broad other medical services category such as eyeglasses, hearing aids, heating pads, incontinence supplies, etc. Average payments for unmatched survey reports of other medical events ($42) were just a small fraction of average payments for matched events ($340) and unmatched Medicare claims ($335) in the same category. This suggests that most unmatched survey events for other medical services are probably not undiscovered matches.

7. Taken together, over 40,000 of the 76,000 unmatched survey events either definitely could not, or very likely would not, match a Medicare bill event record. This leaves 36,000 unmatched survey events to be explained.

8. Estimating conservatively, this means that 52,000 medical events, or 27% of Medicare bill records for community dwelling original sample persons were not reported in survey interviews. (Calculated using 88,000 unmatched Medicare events minus 36,000 possible undiscovered matches among the unmatched survey events)

Unmatched survey events likely to be undiscovered matches

9. On other side, over 16,000 unmatched survey-reported events reported a dollar amount that Medicare paid for the event. These unmatched survey events are very likely to be undiscovered matches.

Ambiguous events

10. This leaves about 20,000 unmatched survey events to be explained. There are many medical services and supplies that Medicare does not cover. For example, physical examinations if the person is well, most alternative medicine services, over the counter remedies, etc. Average payments for unmatched survey events ($64) are considerably less than average payments for matched events ($671) and unmatched Medicare bill record events ($259). These comparatively low average payments suggest two things: many of these events probably are not undiscovered matches, and even if they are, adding them into the file (rather than excluding them as potential duplicates) would not have a significant effect on total payments.

Building A Complete File

MEDICARE COVERED SERVICES

A. A complete file would include all 104,000 matched events. These events which were reported on both the survey and in Medicare bill event records will combine the most accurate and complete information possible from both sources.

B. All Medicare bill record unmatched events (88,000) should also be included. These event records are official records of Medicare program payments and will correct for survey under-reporting.

C. It is more debatable which of the unmatched 76,000 survey records to include. We recommend, and we have included in our file type of service and adjusted file summaries, all unmatched survey reports except the 16,000 records with a Medicare payment. For the reasons discussed above, we assume that these 16,000 records are undiscovered matches that would duplicate some of the 88,000 unmatched Medicare bill event records if they were included.

D. Home health and Hospice records, which were not entered in the event level match, should be added into the file.

TOTAL MEDICAL SERVICES INCLUDING MEDICARE COVERED AND NON-COVERED SERVICES

In addition to A, B, C, and D above, Prescription Drug and Long-Term Facility records should be added to the file.

MISSING PAYMENTS AND PAYERS

This sections describes adjustments made to fill in payment amounts that are missing because the beneficiary did not know how much an event cost, or did not know how the event was paid for (by whom, and how much for each payer). The MCBS staff first established a target reimbursement or total payment for the event, identified all possible sources of payment, and then distributed the total payment across all payers. Missing amounts and payers were filled in using either analytic editing or statistical imputation. This process relied heavily on Medicare administrative records. The guiding principle of retaining as much survey data as possible, and filling in around it, was maintained throughout the process. Where feasible, information about the payers for a specific event, known payment amounts, and target reimbursement were used to determine unknown payment amounts by analytic edits. When insufficient information was available and analytic editing was impossible, unknown payment amounts were completed by statistical imputation.

Different approaches were used with different categories of utilization to define payers and determine payment amounts. Records submitted to the survey/administrative match (which was discussed in the preceding section "Event Level Matching") were handled differently than those, which were not matched. Survey-reported records for dental, medical practitioner, inpatient, outpatient, institutional (other than long term care), and medical equipment and supplies (survey utilization categories DUE, FAE, IPE, IUE, MPE, OPE and PME) were entered into the match with Medicare claims data. After the match, these events were individually assigned target reimbursement amounts, and then source of payment variables and separate payment amounts were calculated for each payer. Other procedures, usually some adaptation of the procedures sketched above, were used to determine payers, target reimbursements and payments for other categories of utilization. In the next sections we discuss how target reimbursements were established, explain the procedures used for matched utilization (the largest category of utilization), and then discuss the smaller and more specific non-matched categories.

Determining target reimbursement One of our primary rules was to establish the target reimbursement for an event with a missing total payment prior to determining or imputing the payment distribution. This was done in a way to establish a target reimbursement that was consistent with payments shown for other similar services in the file. In this way, a credible target reimbursement can be used to inform and control the payment distribution. For Medicare covered services, target reimbursements were developed from Medicare claims because this is a more accurate method than determining the amounts paid by individual sources of payment, and summing them.

Another primary rule was to retain survey-reported payment data, even when it was only partial data, wherever possible. There are situations where retaining the reported payment amounts and establishing the target reimbursement amount without regard to individual payment amounts are mutually exclusive. On a few occasions, the target reimbursement had to be adjusted in order to retain reported payment data.

The rules for establishing target reimbursements depend first on whether or not Medicare claim data are available. If the survey-reported data match a Medicare claim record, or if the Medicare claim record was the only source of information about the service (nothing about the service was reported in the survey), the Medicare claim data were used to establish a target reimbursement. The target reimbursements for nearly 4 out of 5 (79%) of claims submitted to the match were established using Medicare administrative bill payment data.

If the utilization was only reported in the survey (matching to Medicare claims was not successful in identifying a corresponding claims record), the survey data was used to create the target reimbursement. This occurred for about 1 in 5 (21%) of events submitted to the match.

For a small subset of the survey reported events without a matching Medicare claim, but where Medicare was reported as a payer, a different approach was used to create a target reimbursement. A set of regression models, one for each type of event, was developed to predict the target reimbursement from the total charges reported in the survey.

When the respondent did not report a total charge for the event but indicated that Medicaid was a payer, an imputed target reimbursement was created which was consistent with the generally lower payments made by Medicaid.

Filling in Missing Payments and Payers for Matched Utilization Records

The following procedures were used to determine who paid for each event, how much an event cost in total, and how much each payer paid. These procedures were applied to inpatient, outpatient, institutional (other than long-term care), dental, and physician and supplier services, and medical equipment and supplies. These procedures were applied to events in the 1992 Cost and Use file designated: RICIPE (inpatient), RICOPE (outpatient), RICDUE (dental), RICIUE (institutional) and RICMPE (medical and surgical services, equipment and supplies).

Determining Potential Payers Regardless of the method used for imputation, payment amounts were only imputed for potential payers. The total reimbursement for an event was distributed among 11 sources of payment (SOP):

- Medicare fee-for-service
- Medicaid
- Medicare managed care
- Private insurance managed care
- Veterans' Administration
- Employment-based private health insurance
- Individually purchased private health insurance
- Private insurance, source unknown
- Out-of-pocket
- Uncollected liability
- Other public insurance

Out-of-pocket payments are those payments made by the beneficiary or their family, either as cash or through Social Security or SSI checks to a nursing home. Medicare managed care organizations (MCOs) coverage is different enough from fee-for-service coverage to merit its being reported separately. Non-MCO private insurance is characterized as individually purchased or employment-based because there are differences in cost and coverage depending on type. As this information is not known for

residents of nursing homes (the nursing home staff are not likely to know, and thus are not asked, how the insurance was purchased), a third category of private, non-MCO insurance was created for private insurers when the source is not known. Uncollected liability refers to unpaid amounts where there is a legal obligation to pay. If there is an agreement between the provider and a payment source, which reduces the amount the provider can collect for a service, there is no uncollected liability. On the other hand, if the respondent reports a total amount payable and specific payment amounts for all known sources of payment, and the sum of those payments is less than the total amount payable, the difference is considered an uncollected liability. Other public insurance includes Federal or State programs not included in the other categories, such as State drug programs like PACE in Pennsylvania.

An individual's insurance coverage can change during the course of a year. A health insurance time line, created for each person in the 1992 Cost and Use file, provided the basis for determining the potential payers for each event. The time line contained complete insurance information, including Medicare entitlement and enrollment in Medicare MCOs, for every day of the beneficiary's Medicare eligibility during the year. Medicare entitlement and enrollment in Medicare MCOs was captured from CMS administrative data, while information about private insurance was collected in the insurance portion of the interview, and then supplemented by information learned from statements and Medicare claims.

Payer indicators A payer indicator code was used to identify definite and potential payers of the total charge for an event. SOP (Source of Payment) flags were used to initialize the payer indicator. Each SOP flag corresponded to one component of the payer indicator, and could have a value ranging from 0 to 4.

Figure 3. Source of payment (SOP) flag values

0 - Source definitely did not pay
1 - Source definitely did pay, known amount
2 - Source definitely did pay, unknown amount
3 - Source possibly paid, beneficiary was covered at time of event
4 - Source possibly paid, beneficiary may have been covered at time of event

SOP values were set by using survey information about reported events, about the type of provider for the event (that is, whether the service was delivered by a managed care provider or a VA facility), and about the type of insurance coverage and/or program participation. SOP values also depended on Medicare claims data when a survey-reported event corresponded to a Medicare claim (a "matched" event.) Based on all of this information, each source was determined to be a payer, a potential payer, or not a payer of charges for the event.

Payers A source was a definite payer if the SOP for that source had a value of 1 or 2. An SOP value of 1 indicates that the respondent reported that the payer had paid a portion of the charges and also reported a payment amount, or that Medicare claims data provided

that information. An SOP value of 2 means that the respondent reported that a payer paid a portion of the charges, but did not know the exact amount, and no matching Medicare claim was found to provide this information.

Potential payers A source was a potential payer if the corresponding SOP had a value of 3 or 4. An SOP value of 3 meant that either the beneficiary definitely had that type of insurance coverage at the time of the event and the payer may have paid some amount, and/or the beneficiary received the service from that type of payer (i.e., a managed care provider or a VA facility), but did not report it as a payment source. An SOP value of 4 was used when there was doubt about the beneficiary's insurance coverage during the event or about the event date itself.

Non-payers If neither the respondent nor the Medicare claims data indicated that a payer had been a source of payment for an event, the SOP was set to 0.

A more comprehensive discussion of the rules used for setting the SOP flags is included in Technical Appendix D.

Translating payer indicators into sources of payment - A value of 1 for a particular payer indicator meant that the payers paid a portion of the total charge for the event, and a value of 0 meant that the payer did not contribute. Final payer indicator values were determined in one of three ways: 1) directly from the corresponding SOP values; 2) through analytic edits; or, 3) through statistical imputation.

Different rules applied when payer indicator values were set directly from the corresponding SOP values, depending on whether the SOP was determined to be a definite payer, a potential payer, or a non-payer. If the source was a definite payer and the payment amount was known (SOP=1), the corresponding payer indicator was set to 1. If the source was a definite payer but the payment amount was not known (SOP=2), the payer indicator value was set to 1 with one exception: if the event was for dental care or for durable or non-durable medical equipment not usually covered by Medicare, the Medicare payer component was set to 0. The rationale was that if the respondent was not able to report the Medicare payment, then it was more likely that Medicare had not actually paid for the ordinarily non-covered dental services.

When the SOP was a potential payer (SOP=3 or 4), the corresponding payer indicator was set to missing, and imputed (as 0 or 1) in a later step. However, the general rule for setting a payer indicator value based on a corresponding SOP value of 3 or 4 was sometimes modified by the analytic edits, as discussed next.

> NOTE: The Medicare payer indicator value was never set to missing. It was always equal to 0 or 1, unless the SP reported that Medicare had contributed toward the event but did not report the amount and the survey data was not

matched to a Medicare claim. In this case, the SOP value for Medicare was set to 2 and the Medicare delta value was determined as above.

When the SOP was not a payer (SOP=0), the corresponding payer indicator was set to 0, except when the SOP was out-of-pocket or uncollected liability. If the SOP was out-of-pocket or uncollected liability and equal to zero, the payer indicator was set to missing, to be imputed (as 0 or 1) in a later step.

Analytic edits Analytic editing of charge and source of payment data at the event level also determined some payer indicator values. The general goal of the analytic edits was to resolve as many events as possible (i.e., to fully allocate total charges to payers) and to set as many payer indicator values as possible based on logic and knowledge of payer policies. The edits resolved some events without using a hotdeck procedure to impute payment sources or amounts.

The analytic edits relied on having both unambiguous SOP values and external information about interaction among the insurance or payment sources. Edits for three of the payment sources (Medicaid, MCOs, and VA) depended on information specific to those payers, but payer indicators for other payment sources were also affected. The analytic edits are discussed fully in Technical Appendix C, as they apply to each source of payment.

Medicaid: Analytic edits were used extensively when Medicaid was a potential or actual source of payment for an event. One set of edits--designed to reflect the role of Medicaid as the payer of last resort--ensured that Medicaid could not be a payer if payments were reported or imputed for another third-party insurer (except Medicare), or if the provider was an MCO or VA facility. Another set of edits was developed for dual Medicaid/Medicare eligible beneficiaries whose cost-sharing liability is covered by Medicaid.

Private and Medicare MCOs: Managed care organizations (especially Medicare-contracting MCOs) often operate differently than other third-party payers and tend to have unique payment patterns. For instance, risk and (to a lesser extent) cost Medicare MCOs are paid a set fee per enrolled Medicare beneficiary (called a capitated amount) designed to compensate the MCO for the expected costs of delivering Medicare's package of benefits. There are no Medicare claims or Medicare or insurance statements indicating the total charge for events covered by the capitated amount. Often the respondent only knows the copay amount, if there was one. Also, MCOs often provide "Medigap"-type coverage by paying for most of the member's deductibles and copays for Medicare-covered benefits. A beneficiary who belongs to an MCO does not need private Medigap insurance or Medicaid coverage for these amounts. Thus, payment patterns for MCO beneficiaries tend to be simpler than those for fee-for-service beneficiaries. The set of analytic edits for MCOs attempted to account for these simplified patterns and for

the respondent's usual inability to report charges and payments for events. The MCO edits also attempted to avoid creating "illogical" payment patterns.

Veterans' Administration (VA) coverage: If VA was a payer, no uncollected liability amounts were allowed. As both the insurer and provider of services, the VA does not "charge" more than it will be reimbursed by other payers. In this respect, services provided by the VA are similar to those provided by MCOs.

General Edits: At the beginning of the analytic editing, and after each main section of edits, an attempt was made to resolve events through addition or subtraction. Events without a known total charge but with a complete payer indicator vector (i.e., each payer was identified as either having paid or not paid for an event and each payer's amount was known) were completed by summing across all payment sources to derive the total charge. Events with a known total charge and complete except for one missing payment amount or payment source were completed by subtraction. The excess of charges over known payment amounts was attributed to the known payer, or the one missing payer indicator was set to 1 and the excess allocated to that payer.

If a service was provided free of charge, all payer indicators and payment amounts were set to 0. However, if the respondent reported an event as free, but also reported that a source other than Medicare or Medicaid paid something for the event, the total charge was reset to "missing", and imputed.

If a source was a potential payer for an event, or if the respondent reported that the payer had contributed to an event but did not know the amount, it was assumed that the payer was not actually a source if the current sum of reported payments equaled the reported total charge.

Payer Indicator Imputation

Delta components that still had missing values after accounting for survey data, Medicare claims data, and the analytic edits were imputed through a hotdeck procedure. The hotdeck procedure used completed payer indicators by identifying similar cases that served as donors for comparable cases with incomplete vectors (beggars). Comparability was usually defined in broad terms so that there were multiple choices for each event that needed payer indicator imputations.

If Medicaid was a payer, a Medicaid payment amount was calculated as a percentage of coinsurance and deductible for the Medicare service.

Other Utilization (Not Matched)

The following procedures were used to determine who paid for each event, how much an event cost altogether, and how much each payer paid, for events that we did not attempt to match to Medicare claims data on a service-by-service basis. These procedures were applied to home health and hospice services. (The procedures used for missing payments or payers for prescription drugs and facility utilization are described separately. We thought it would be more helpful to readers if we kept all the information on how the long term facility and prescription drug records were created, edited, and had missing data filled in one complete section. For information on the editing and creation of these types of utilization, refer to the Prescription Drugs and Long-Term Facility segments in Section 4, Edits). Long term facility and prescription drug utilization are presented in the 1992 Cost and Use file as event-level records designated: RICFAE (facility) and RICPME (prescription medicines). Hospice and home health records are presented as summary records only.

Hospice Services

Hospice utilization is unusual in terms of Medicare administrative records because it is the only utilization that is recorded in two different ways, in two different files. The beginning and ending dates of the hospice benefit periods are recorded in the enrollment database (EDB), while the bill records are part of the national claims history repository (NCH). This dual reporting served as an internal check on the dates of service on the billing records.

Determining and imputing payment amounts With a target reimbursement amount (this represents the "total cost" of the event), and delta values indicating which payers contributed some payment toward the total, the share "amounts" paid by the individual payers could be determined.

If Medicare payments were known to be incomplete, and utilization for the missing periods was completed by editing from the existing billing records. The hospice benefit is paid on a per-diem basis, and the missing data were completed with average per diem rates calculated from existing bills. Virtually all services provided to the hospice beneficiary are fully covered by Medicare, and as there are no co-payments or deductibles, there is no cost sharing (prescribed medicines are an exception, as there may be a small co-payment for drugs, which are reported separately, and also inpatient respite care for which the patient pays 5% of the Medicare allowed rate - under $5 in 1993). Hence, the Medicare reimbursement is the target reimbursement, and Medicare is the sole payer of hospice bills. Hospice bills were not matched; as a result, there is some overlap between hospice utilization and events reported in the survey. The overlapping survey events are usually, but not always, home health events.

Home health

The home health use and payment records in the Cost and Use file are designed to represent events where medical care, as opposed to personal care and support, was furnished to the sample person. The decision to include only medical services in the user file in no way derogates the importance of unpaid assistance in maintaining the health and well being of Medicare beneficiaries. It simply reflects the primary emphasis of the MCBS Cost and Use file, which focuses on use of, and payment for, formal medical care services.

Home health events, like prescription drug events, are undated on the survey. For reference periods that spanned two years, the first step was to allocate services proportionately into 1992. The rules used to do this were identical with the procedures in the ADJUSTING FOR MISSING DATES AND UNDATED SERVICES discussion below near the end of Section 5. At this stage, a home health "event" could have represented one or more home health visits. Bundled events with multiple visits were unbundled for the allocation of home health services across years. (Note, however, that home health use and costs are summarized at the type of service and person levels in this file, and home health "event" level data is not shown. The summaries do contain counts of home health visits.)

Survey event records were originally classified in the interview according to whether a professional or friend provided the home health services. This distinction was used in separating out home health services that were not medical in nature. In winnowing down the file to medical services only, the following decision rules were used to EXCLUDE non-medical home health services:

1. Exclude services provided by a friend where the out-of-pocket payment, if any, was equal to the total charge for the service. (The reasoning is that even if the friend was paid for delivering a service, it was very likely non-medical in nature if there was no other payer).

2. Exclude services provided by a professional where the out-of-pocket payment was equal to the total charge for the service AND the person answered NO to the question whether the professional gives nursing/medical treatment.

3. Exclude all housekeeping/cleaning services unless Medicaid is listed as a payer.

4. Exclude all "meals-on-wheels" types of services.

After these allocation and exclusion operations, the remaining survey reported medical home health services were matched (not at the event level but at the person level only) to Medicare bills for home health services. The survey reports and Medicare bills were

combined to provide the most accurate and complete summary possible of number of visits and payments (broken down by source of payment such as Medicare, out-of-pocket, etc.)

Other non-covered utilization

For services not covered by Medicare, we made an estimate of medical usage during periods not covered by interviews, in order to produce a file that can be used to estimate full expenditures for the year. For periods of missing data, we first determined the use of services not covered by Medicare, by determining the number of events of the type per day for the covered period, multiplied that number by the "gap" days, and added the number of events to the total known events of the same type. Likewise, to get the adjusted sums for all payers, we calculated the costs per event per payer per day, and then multiplied that figure by the adjusted number of events within payer. If the beneficiary had no interview data about the use of medical care, we used averages from a donor--a respondent who had characteristics in common with the beneficiary with missing data.

> NOTE: These adjustments are person-level adjustments, only, and are not reflected in the event records. In addition, they only cover Medicare covered services. There is no adjustment for non-covered services.

Prescription Drug Data

The approach used to fill in missing drug payment data was similar to that used for other missing payment amounts described above. The first step was to establish a total payment amount for each drug event. First preference was given to using survey reports of the total payment for the drug. In 55% of drug events on the file, the total payment was a survey report. For 30% of drug events, an administrative drug-pricing source (National Drug Data File User Manual published by First Data Bank - "The Blue Book") was used to impute prices. The administrative source was used only when no total payment was reported, and it was never used to supersede the survey reported payment. Finally, 15% of drug events had total payments established using statistical imputation techniques.

After the total payment was established for each drug event, potential sources of payment were identified using a similar approach to that outlined earlier in Section 5. In the last step, the total payment amount was distributed across the sources of payment. In the 85% of cases where a total payment was available from either a survey report or the "Blue Book", unknown payment amounts for a specific payer were handled by accounting techniques and analytic edits before employing statistical imputation. In the 15% of cases where the total payment was derived by statistical imputation, the payer amounts were also derived through statistical imputation.

Preparation of survey reported data Prior to imputation for costs and payments, the prescription drug data collected in the survey were edited for consistent spelling. Although respondents are encouraged to save empty packaging from all prescription medicines, inconsistencies in spelling are sometimes introduced as the data are collected. As a first step in processing the prescription drug data, MCBS staff edited the records to ensure that the same drug was always reported in the same way. All unique drug name spellings supplied in the survey from Rounds 2 through 5, including both community and facility survey responses, were gathered together in a single list. Using the 1992 Blue Book, MCBS staff manually assigned corrected spellings to each unique supplied spelling.

Preparation of Blue Book data The 1992 Drug Data Bank File User Manual from First Data Bank ("Blue Book") served as a pricing reference and as a source of therapeutic class for prescription medicines. However, survey reports of total payments were given preference over a "Blue Book" price because we could not match MCBS records and "Blue Book" records exactly on all fields. The Blue Book generally identifies the name, form, strength, and packaging size of the drug in a single entry. The MCBS collected prescription size in the survey, but could not collect the packaging size of the drug prescribed. In the survey, form and strength are also collected, but as separate items, not as part of the name. In the initial match, therefore, a Blue Book name "Septra DS Tab 800 mg" had to be changed to "Septra DS", to increase the likelihood of a match between the two sources on name.

Assignment of wholesale prices - In the Blue Book, a wholesale price is assigned to each National Drug Code (NDC) entry. The NDC is an 11-digit code; the first nine digits identify a drug (including form and strength), and the last two digits identify the packaging size. As noted above, the MCBS does not collect packaging size, but prescription size, and unit average wholesale prices can differ substantially by packaging size. Using a relative frequency distribution of packaging sizes within each drug type, weighted by utilization rates from CMS's Medicaid Statistical Information System, MCBS staff developed a composite price for drugs that come in multiple package sizes.

After both survey data and Blue Book data were cleaned as described, survey prescription data were matched to the modified Blue Book information by drug name, form, strength and packaging size, in that order, to develop an average wholesale price. Often, we were not able to match on all four variables. If the survey drug name was not known or could not be matched, an average wholesale price was imputed. If the drug name was known but form or strength were not known, the missing characteristic was imputed and the average wholesale price was then obtained through a match to the Blue Book. For example, if the respondent reported a prescription of Diamox but did not know the strength, an average wholesale price was imputed using the weighted average price of all Diamox prescriptions (developed using a frequency distribution of drugs by National Drug Code in the Medicare-Medicaid dual eligible population).

A small number of survey entries could not be translated to any drug listed in the "Blue Book". In general, these entries were either misspellings which made it impossible to determine the drug name or not really even a specific drug (e.g. "little green pills"). These entries were classed as "untranslatable", and an average price was imputed based on frequency distributions of drugs taken by the Medicare and Medicaid dually eligible population.

In some cases the size of the prescription was known but the price was not. Average unit costs (per pill, per milliliter, etc.) were then multiplied by the prescription size, to derive a whole prescription cost. In other cases, prescription size was estimated through the respondent's answers to a series of probe questions, which were asked during the interview when the respondent did not know the size of the prescription.

Converting average wholesale price into event price Establishing a price for prescription drug records with no survey reported price began with the assignment of an average wholesale price. Event prices that were less than $1.50 were reset to missing, and imputed statistically. Non-missing wholesale prices were multiplied by a pricing factor, which varied depending on the likely payer(s) of the event. Five pricing factors were developed: retail, managed care organization, VA, Medicaid and other public insurer. The retail pricing factor was actually a series of factors which reflected empirical evidence of the relationship between the average wholesale price and what the respondents reported paying. The retail factor was 250 percent, 121 percent, or 90 percent, depending on the wholesale price of the drug. The managed care pricing factor was based upon a General Accounting Office study of prescription drug data for 1991 from managed care organizations, where it was reported that managed care organizations pay approximately 70 percent of the average wholesale price of prescription medicines and have an average dispensing fee of $2.27. The VA factor was developed using VA drug cost data, which was provided by the Department of Veteran's Affairs. The Medicaid pricing factor was developed using composite data from HCFA's Medicaid Drug Rebate System, and included a dispensing fee, a discount of the average wholesale price (89 percent) and a rebate percentage of 21 percent.

Determining target reimbursement Target reimbursements were developed differently for prescription medicines than for other services. (Target reimbursements for other types of services are described above in Section 5). Generally, Medicare does not cover prescriptions, and therefore there were no Medicare claims for price comparisons. In place of the unavailable Medicare claims data, adjusted "event prices" (described above) were used to develop target reimbursements.

The target reimbursement is defined as the price that the beneficiary paid for a single purchase of a single drug. For single purchases (one unique medicine, purchased only once and not refilled), the price reported by the respondent was the target reimbursement. If the respondent could not give a price, the event price, adjusted by the appropriate

pricing factor (discussed below) was the target reimbursement. For multiple purchases (a single prescription, filled multiple times, or multiple prescriptions), the target reimbursement was developed as for single purchases, and then divided by the number of purchases to yield a target reimbursement for each purchase.

If several drugs were reported together ("bundled"), but the total cost was not known, a target reimbursement was developed for each drug in the bundle, based upon the event price adjusted by the appropriate pricing factor. If several drugs were bundled together and a total cost was reported, that total cost was used to control the imputation of the individual drug prices. A relative percentage of the total cost was developed for each drug, using the event price adjusted by the appropriate pricing factor, then those percentages were applied to the reported total cost and the result became the target reimbursement for each drug. If the event price for one or more of the drugs in the bundle was missing, an average price for all strengths and forms of the drug was used in the computation, unless the drug name was not known, in which case an average event price (computed across all drugs, about $30) was used. These averages were then used to calculate relative percentages, which were then applied to the amount reported in the survey for the bundle.

Determining potential payers Potential payers for prescription medicines were determined in essentially the same way that potential payers were identified for matched utilization, as described above in Section 5.

Post-imputation checks In line with our overall approach, survey data were retained unless strong evidence suggested that they were wrong. After statistical imputation, it was occasionally necessary to change survey reported target reimbursements for drug events.

Sometimes, a sample person purchased a drug through a public program or through a managed care organization and reported that the out-of-pocket expense was the "total cost" of the drug. Following the procedures outlined above, the out-of-pocket cost would become the target reimbursement for the event. In order for the target reimbursement to be changed, all of the following had to be true:

> The source of payment flags for Medicaid, VA, other public insurer, or a managed care organization, were coded 3 or 4, indicating these payers could have paid for the event, even though they were not so identified by the respondent.
>
> The event price was not generated using any imputed information on form, strength or volume.

The target reimbursement was less than 50 percent of the average wholesale price adjusted by the appropriate pricing factor.

The out-of-pocket amount reported was equal to the target reimbursement, was less than $10, and was divisible by $0.25.

When all of these conditions were met, the target reimbursement was changed to the average wholesale price, adjusted by the appropriate pricing factor. If this pattern was observed in the total price of a bundle, it was assumed that all drugs in the bundle were reported incorrectly, and all target reimbursements were changed. In all cases, the reported out-of-pocket expenditure was retained.

If the same situation applied to an event where one of the payers was private insurance, the rules for changing the target reimbursement were not as stringent. If the source of payment flags indicated that the beneficiary's private health insurance could have covered the drug purchase, and the respondent said that the out-of-pocket expenditure was the total cost, the target reimbursement was changed to the event price adjusted by a pricing factor. All drugs reported as a bundle were treated the same way. Out-of-pocket amounts were retained as reported.

Special cases After statistical imputation, 32 sample people had negative aggregate managed care payments for drugs, and 5 people had negative aggregate VA payments for drugs. Negative dollar amounts occur in imputation because for a given prescription, the out-of-pocket payment might be higher than the actual cost of the drug. For example, the cost of a 10-day supply of Ampicillin will probably be less than a $5-dollar copayment. In some cases, however, negative prices were the result of an incorrect distribution of out-of-pocket costs when the total charge of a bundle of prescriptions was missing. Consequently, we decided to cap out-of-pocket payments for these two groups of people. After capping payments, 16 sample people still have negative aggregate managed care drug payments, and 2 people have negative aggregate VA drug payments.

ADJUSTMENTS FOR MISSING DAYS AND UNDATED USE

This section describes the adjustments made (at the person level, not the event level) to:

1. compensate for data that are missing because some periods of the beneficiary's Medicare entitlement were not covered by interviews. CMS administrative records are used to establish the exact period of Medicare entitlement during 1992 and calculating the number of Medicare days;

2. allocate undated survey events, primarily prescription drugs and other medical equipment, between years where interview reference periods spanned two years.

Calculating Medicare covered days It is important to define, for each beneficiary in the sample, the exact period of Medicare entitlement during 1992. It is also important to accurately count the number of days in each setting for persons living in the community and living in long term care facilities.

For most sample persons, the period covered by the survey and the period of the beneficiary's Medicare entitlement are identical: they both cover all 366 days of 1992. There are, however, exceptions where the survey period and the entitlement period do not coincide exactly. Differences between the survey and Medicare entitlement dates fall into two categories: the survey period is greater than the Medicare entitlement period; or, the survey period falls short of the Medicare entitlement period.

In a few cases, the date of death recorded in the survey does not agree with the date of death in CMS records. In these cases, the date of death collected in the survey appears as the latest boundary for Medicare covered days, unless CMS *billing data* indicated that the date of death occurred after the survey reported date of death.

There are 11 people in the file who were not entitled to Medicare in 1992, even though some of them participated in the survey. Three of these people retroactively lost entitlement; the other eight died before they reached their entitlement start date, but after they were selected for the MCBS sample (entitlement dates can be recorded as much as 6 months in advance). Records for these people show 0 (zero) Medicare covered days.

The Medicare entitlement period is longer than the period covered by the survey when a Round 1 individual left the survey before the end of the year, or died without naming a proxy respondent. This is also true for people who were never interviewed about their use of services in 1992 - the entire Round 4 and Round 7 supplemental samples. The most common reason for incomplete data is the beneficiary's refusal to participate further in the survey. If the beneficiary participated in the survey for at least 60 percent of the period they were eligible for Medicare during the year, the sample person was retained for the 1992 Cost and Use file. If the beneficiary left the survey earlier, that is, the

interviews covered less than 60 percent of this sample person's eligibility in 1992, the beneficiary and the survey data were not retained.

When a sample beneficiary dies or otherwise terminates entitlement to Medicare, a closing interview is conducted with a proxy, or with nursing home staff if the beneficiary is institutionalized. In this way, the survey is designed to capture complete information about people who die or lose entitlement before the end of the year. In a few cases, the beneficiary cooperated with the interviewers for most of the year but died without naming a proxy, leaving unreported the period of time between the last interview and the beneficiary's death. In these cases, as with the beneficiaries who "dropped out" of the survey and the supplemental samples, we used what the beneficiary reported during interviews and Medicare billing data (which is known) to guide the imputation of non-covered services (which are not known) to fill in the gaps in reporting.

Calculating community days and facility days The MCBS sample includes people who are institutionalized as well as those who live in the community, and follows people as they make the transition from one type of living situation to the other. For purposes of analysis, it is important to be able to identify people in either situation, and for people who made a transition during the year, to be able to place them in one category or the other for the appropriate amount of time. We provide three variables to show a person's status in this regard: total number of days entitled to Medicare; number of days where the beneficiary was living in the community; and number of days where the beneficiary was living in a facility.

Group I - Information about the community/facility status for this group was collected in each interview in 1992. This is the only group that will ever show a transition from community to facility, or vice versa.

Groups II and III - For beneficiaries in the supplemental samples, we deemed the entire period of Medicare entitlement to be in the same situation as we found them at the initial interview. If a beneficiary was in the community when initially interviewed in Round 4 or Round 7, the beneficiary was deemed to have been living in the community for the entire Medicare entitlement period. Similar logic applies for residents of facilities.

Group IV - These beneficiaries were never interviewed, so information about their living situation was imputed from a donor population. If the donor was living in the community during 1992, the recipient was deemed to have been living in the community for the entire Medicare entitlement period. Similar logic applies for residents of facilities.

Once the periods of Medicare entitlement and living situations are established, utilization reported in the survey is validated by and, in many cases, supplemented by information

reported on claims and bills from CMS's national claims history database. This is accomplished by the matching survey-reported utilization to the CMS records that was described earlier in Section 5.

Allocating services between years The cost and use data collected during the interviews collecting 1992 data (that is, Rounds 2 - 5) cover more than just that calendar year. Each interview serves as a boundary to the next interview - the beneficiary describes medical care that took place "since the last interview" - and those boundaries are generally not the beginning or ending of the calendar year. As a result, the first (Round 2) or last two (that is, Rounds 4 and 5) interviews generally include utilization that covers part of two calendar years. To adjust the utilization in these cases, dated event records were edited to remove those that took place outside of 1992, and undated events were pro-rated according to the number of 1992 days in the interview reference period to total days in the reference period.

Simply pro-rating use between the two calendar years was considered, but rejected. By assuming that use occurred in both years, this procedure could overstate the number and rate of persons using services in a year. In place of this, a random number generator was used to assign services (primarily prescription drugs and other medical events) to calendar years. The probability of an event being placed in 1992 was based upon the ratio of 1992 days in the reference period to total days in the reference period. For example, assume a reference period had 120 days and 90 of these days were in 1992. For each event, a random number between 1 and 120 was generated. For all events where the random number was 90 or less, the service was allocated to 1992. For all events with random numbers between 91 to 120, the service was allocated to the other year. This allocation process reduced the number of prescription drug users from a potential 8,990 to 8,802 in 1992. The number of total reported drug events, 192,274, was reduced to 181,232 allocated to 1992.

Filling in Medicare covered days not surveyed When there is a gap in survey data, that is, a period for which a sample person was enrolled in Medicare but was not covered by a survey interview, it is necessary to estimate the medical service usage during that gap period. For persons with gaps who were interviewed in 1992, reported services were simply prorated upward to cover the gap. For example, for prescription drugs the number of prescriptions per day were calculated for the interview period, and multiplied by the number of gap days. This assumes, in effect, that the person used prescriptions at the same rate in the interview and gap periods. Likewise, to get adjusted sums for all payers, the cost per prescription per payer per day was calculated, and multiplied by the adjusted number of prescriptions for each payer.

If the sample person was not interviewed (e.g. supplemental sample persons), a different approach was used. To cover these non-interview gap periods, a donor was selected who was similar to the person in terms of personal and economic characteristics. The donor's

use of prescription drugs (measured in prescriptions per day and cost per prescription per payer per day) was used to impute use and payment data.

Sample Design and Estimation

This section opens with a brief description of the sample design (also discussed in the Introduction), the population actually covered by the original sample (persons enrolled as of January 1, 1991), and the survey operational considerations which led to the use of the January 1 population (modified to represent the "always-enrolled") for earlier MCBS public use file releases having to do with the issue of access to care.

Next, follows a restatement of the purpose of the 1992 Cost and Use file. That purpose is related to a particular view of the Medicare population, namely, beneficiaries ever-enrolled during calendar year 1992. Adjustments to the data for the original sample to account for individuals in the target population for the 1992 Cost and Use file but not represented in the surveyed population are discussed. Various "views" of the 1992 Medicare population (always-enrolled, ever-enrolled, and midpoint) are presented for comparison purposes.

Following the comparison is a general review of person level response rates by panel. Guidelines for preparing population estimates using full sample weights and variance estimates using replicate weights are then reviewed.

Sample Design

The MCBS is a continuous, multi-purpose panel survey of Medicare beneficiaries. The target population of the study consists of aged and disabled persons enrolled in one or both parts of the Medicare program, that is, Part A (Hospital Insurance) or Part B (Supplementary Medical Insurance), and residing in households or in long-term care facilities in the 50 States, the District of Columbia, and Puerto Rico. The sample design is a stratified area probability design with three stages of selection: (1) selection of 107 primary sampling units (PSUs), which are metropolitan statistical areas and clusters of non-metropolitan counties; (2) selection of ZIP code clusters within the sample PSUs; (3) selection of Medicare beneficiaries within the sample ZIP code clusters. The sample was designed to yield complete annual health care cost and use data on 12,000 beneficiaries.

Initial interview questionnaire

This baseline questionnaire is used for the first interview when a sample person is added to the survey, that is, Round 1 for the original sample, Round 4 for the 1992 supplement, Round 7 for the 1993 supplement, Round 10 for the 1994 supplement, etc. In the initial interview, we collect information about the national origin, age, education and income of the sample person. The interviewer also verifies the sample person's address and telephone number and obtains the names and addresses of people who might be willing to serve as proxy respondents. The interviewer also uses this opportunity to acquaint the respondent with the intent of the survey and to familiarize him or her with the MCBS calendar, and to emphasize the importance of keeping accurate records of medical care and expenses.

In subsequent interviews, some of the information collected in the initial interview is updated. For example, the sample person's designation of his or her race is not likely to change, and will not be asked about again. On the other hand, the sample person's address or telephone number may change, so this information is verified in every interview, and updated when necessary.

Core questionnaire (community)

The core questionnaire is the major component of the community instrument. The questions focus on the use of medical services and the resulting costs, and are asked in essentially the same way each and every time the sample person is interviewed (after the first time). In each interview, the sample person is asked about new encounters, and to complete any partial information that was collected in the last interview. For example, the sample person may mention a doctor visit during the "utilization" part of the interview. In the "cost" section, the interviewer will ask if the sample person has any receipts or statements from the visit. If the answer is "yes", the interviewer will record information about costs from the statements, but if the answer is "no," the question will be stored until the next interview.

Supplement to the core questionnaire (community)

Supplemental questions are added to the core questionnaire to gather information about specific topics. The Round 4 supplement focuses on health status and access to care. It includes questions about the sample persons' general health (including standard measures such as IADLs and ADLs), their sources of medical care, and their satisfaction with that care.

Interviewer remarks questionnaire

This questionnaire is completed by the interviewer after every interview with the sample person. The interviewer is asked to evaluate the sample person's ability to respond to the questionnaire, and to provide some information about the interview (for example, if the questionnaire was answered by proxy, the interviewer provides reasons why the proxy was necessary). The interviewer is also encouraged to provide comments that will assist the interviewer in remembering unique facts about the sample person, such as hearing or vision impairments, or that the sample person cannot read.

Facility Questionnaire

The facility questionnaire is conducted conventionally using pen and paper in the facility where the respondent is residing at the time of the interview. Information is obtained from facility records; therefore, the beneficiary is never interviewed directly. It was decided early in the design of the MCBS not to attempt interviews with sample persons in facilities, or with their family members. For that reason, the facility questionnaires do not ask about attitudes or other subjective items.

If an institutionalized person returns to the community, a community interview is conducted. If the sample person spent part of the reference period in the community and part in an institution, then a separate interview is conducted for each period of time. In this way, a beneficiary is followed in and out of facilities and a continuous record is maintained regardless of the location of the respondent.

Components of the Facility Questionnaire

The facility instrument consists of the following components:

Facility eligibility screener
Initial (baseline) questionnaire
Core questionnaire
Supplement to the core questionnaire

Facility eligibility screener

This questionnaire gathers information about the facility to determine the facility type. The initial interview is conducted with the facility administrator. All other interviews are conducted with the staff designated by the director. A facility screener is administered upon the sample person's admission to a new facility, and once a year thereafter (in Rounds 4, 7, and 10) to capture any changes in the facility's size or composition. The screener is not administered if the sample person simply re-enters the same facility.

Initial (baseline) questionnaire (facility)

This questionnaire gathers information on the health status, insurance coverage, residence history and demographics of the sample person. This questionnaire is administered the first time the sample person is admitted to a facility.

Core questionnaire (facility)

This questionnaire parallels the core questionnaire for the community, collecting information about use of medical services and their associated costs, including the facility cost. Like its community counterpart, this questionnaire is administered in each and every interview after the first one, as long as the sample person continues to reside in the facility.

Supplement to the core questionnaire (facility)

This questionnaire is asked once a year (in Rounds 4, 7, 10) to update our information about the sample person's health status. It includes questions about the sample person's general health (including standard measures such as IADLs and ADLs), but excludes the questions about access and the subjective questions about satisfaction with care.

Table 5.1 - Components of the Community Questionnaire

UPD	NAME/ADDRESS UPDATE
IN	INTRODUCTION
ENS*	ENUMERATION
EN	ENUMERATION
HI	HEALTH INSURANCE
UTS*	UTILIZATION SUMMARY
DU	DENTAL UTILIZATION AND EVENTS
ER	EMERGENCY ROOM UTILIZATION AND EVENTS
IP	INPATIENT HOSPITAL UTILIZATION AND EVENTS
IU	INSTITUTIONAL UTILIZATION
OP	OUTPATIENT HOSPITAL UTILIZATION AND EVENTS
HHS*	HOME HEALTH UTILIZATION SUMMARY
HH	HOME HEALTH UTILIZATION AND EVENTS
MP	MEDICAL PROVIDER UTILIZATION AND EVENTS
OM	OTHER MEDICAL EXPENSES UTILIZATION
PMS*	PRESCRIBED MEDICINE SUMMARY
PM	PRESCRIBED MEDICINE UTILIZATION
ST	CHARGE QUESTIONS (STATEMENT SERIES)
NS	CHARGE QUESTIONS (NO STATEMENT SERIES)
CPS*	CHARGE/PAYMENT SUMMARY
AC**	PROVIDER PROBES/ACCESS TO CARE
SC**	SATISFACTION WITH CARE
UC**	USUAL SOURCE OF CARE
HS	HEALTH STATUS AND FUNCTIONING
US	USUAL SOURCE OF CARE
DI	DEMOGRAPHICS/INCOME
CL	CLOSING MATERIALS
IR	INTERVIEWER REMARKS

* Summary sections - Updates and corrections are collected through the summaries. The respondent is handed a hard copy summary of information gathered in previous interviews, and is asked to verify the material. Changes are recorded if the respondent notices information that is not accurate.

** The data collected in these sections is not included in this public use file. These data appear in the Access to Care series.

Table 5.2 - Components of the Facility Questionnaire

NOTE: This release contains information from all sections

Facility Eligibility Screener

FQ Facility questions

Initial interview (facility)

A Demographic/Income
B Residence History
C Health Status and Functioning
D Health Insurance
L Tracing and Closing

Core questionnaire (facility)

A Residence History
B Provider Probes
C Medicine Summary
D Inpatient Hospital Stays
E Medical Charges
F Tracing and Closing

Supplement to the core (facility)

C Health Status and Functioning
D Health Insurance

Summary Counts

The Codebook in Section 2 provides un-weighted frequency counts of categorical variables, which analysts can use to check tabulations of these variables. The Codebook does not contain similar information for continous variables, such as cost and payment amounts. The table of weighted summary counts below is intended to allow analysts to benchmark their tabulations of MCBS payment variables. The table is created from the adjusted payment amounts from the Service Summary (RIC SS) and weighted by the person weight (CS1YRWGT) from the RIC X. The payments represent 37 million Medicare beneficiaries. **All payment amounts are in thousands.**

Service	Total	Medicare	Medicaid	Medicare HMO
Dental	$ 5,172,012	$ 4,999	$ 106,255	$ 43,515
Facility	$ 54,051,414	$ 554,112	$ 28,852,015	$ -
Home Health	$ 8,622,836	$ 7,695,585	$ 88,589	$ 2,089
Hospice	$ 808,204	$ 807,992	$ -	$ -
Inpatient Hospital	$ 78,865,973	$ 65,737,851	$ 1,165,661	$ 2,386,131
Institutional (SNF)	$ 4,000,962	$ 2,693,890	$ 302,166	$ 76,878
Medical Provider	$ 56,870,093	$ 33,890,073	$ 1,641,019	$ 1,108,779
Outpatient Hospital	$ 19,255,355	$ 11,148,533	$ 753,070	$ 522,546
Prescribed Medicines	$ 16,636,842	$ 48,466	$ 1,663,032	$ 110,058
Total	**$ 244,283,691**	**$ 122,581,501**	**$ 34,571,807**	**$ 4,249,996**

Service	Private HMO	PHI-Employer	PHI-Individual	PHI-Unknown
Dental	$ 40,370	$ 479,212	$ 60,127	$ -
Facility	$ -	$ -	$ -	$ 348,437
Home Health	$ 628	$ 86,914	$ 21,971	$ -
Hospice	$ -	$ -	$ -	$ -
Inpatient Hospital	$ 251,778	$ 3,659,549	$ 2,091,138	$ 58,798
Institutional (SNF)	$ 35,365	$ 140,161	$ 186,537	$ 379,362
Medical Provider	$ 633,197	$ 4,471,944	$ 3,333,896	$ 91,927
Outpatient Hospital	$ 274,719	$ 1,967,565	$ 1,571,788	$ 100,897
Prescribed Medicines	$ 461,320	$ 3,177,836	$ 490,717	$ -
Total	**$ 1,697,377**	**$ 13,983,181**	**$ 7,756,174**	**$ 979,421**

Service	VA	Out of Pocket	Uncollected	Other
Dental	$ 53,922	$ 4,047,933	$ 246,951	$ 88,726
Facility	$ 720,452	$ 21,097,541	$ -	$ 2,478,856
Home Health	$ 768	$ 535,010	$ 1,152	$ 190,130
Hospice	$ -	$ 212	$ -	$ -
Inpatient Hospital	$ 1,128,918	$ 1,565,308	$ 269,516	$ 551,326
Institutional (SNF)	$ 3,379	$ 101,948	$ 69,571	$ 11,716
Medical Provider	$ 111,526	$ 10,205,088	$ 893,977	$ 488,667
Outpatient Hospital	$ 418,770	$ 1,857,415	$ 261,576	$ 378,475
Prescribed Medicines	$ 192,265	$ 9,328,538	$ 299,837	$ 864,773
Total	**$ 2,630,000**	**$ 48,738,993**	**$ 2,042,580**	**$ 5,052,669**

IMPUTATION OF MEDICAL COST AND PAYMENT DATA

Amy M. England[1], Katie A. Hubbell[1], David R. Judkins[1], Svetlana Ryaboy[1]
Westat, Inc., 1650 Research Blvd., Rockville, MD 20850

Keywords: Gibbs Sampling, Hot Deck Imputation, Compositional Data[1]

Medical cost and payment data are the primary focus of the Medicare Current Beneficiary Survey (MCBS). These data are compositional data (data where a finite series of random variables are non-negative and sum to another random variable). There is a large variety of missing patterns that are neither nested nor ignorable. A paper from last year presented a new technique for creating a complete set of compositional data while preserving all partial data and maintaining many types of consistency. This year, we present the results of applying the method to actual MCBS data on prescription drugs. Since the method is known to be extremely CPU intensive, a primary point of interest will be the feasibility of applying the method to a dataset with about 245,000 records and nine possible payment sources.

1. Introduction

The imputation of costs and payment sources for prescription medicines is a critical area for the Medicare Current Beneficiary Survey (MCBS) given the ongoing national debate about whether to expand Medicare coverage to include prescription medicines. There were a substantial number of partially complete reports about purchases of containers of prescription medicine. One solution is to impute the cost where necessary, discard partial payment data, and impute whole payment vectors as proportions to be applied to the cost. This solution was used for example on the 1987 National Medical Expenditure Survey (Hahn and Lefkowitz, 1992, p22). Judkins, Hubbell and England (1993), presented an alternate solution that allows the retention of all partial data payment and cost data. They presented an evaluation of the algorithm on an artificial example. That evaluation focused on the ability of the algorithm to minimize nonresponse bias. In this paper, we evaluate the algorithm in terms of practicality by presenting the results of its application to the 245,000 records for individual containers of prescription medicine in the 1992 MCBS.

In the following sections, we review briefly how prescription drug data are collected in the MCBS, define some notation, present some information on the patterns of missingness observed in MCBS prescription data, review the algorithm (some improvements have been made over the version presented last year), and, finally, present results and ideas for future improvements.

2. Data Collection

The MCBS has a modified panel design where a core panel is supplemented once a year with new additions to the eligible universe and additional beneficiaries from the original cohort so as to maintain cross-sectional precision despite deaths and attrition in the panel. Interviews are conducted roughly every four months. The reference period for each interview extends from the date of the prior interview to the date of current interview. Data are collected about the utilization of health care services, the costs of these services, and expenditures (personal and third-party) for these services.

MCBS data are collected by CAPI (computer assisted personal interview). Interviewers carry laptop computers into the homes of Medicare beneficiaries and run a program that guides them through the interview. Figure 1 mimics a typical screen for collecting information about payments for a health care event after the cost has been determined. Figure 2 shows how it might look after completion. Note that the program presents a list of possible payment sources for the event and that the list is tailored to the beneficiary's insurance status and program participation. The payment sources mentioned by respondents were grouped into the nine categories shown in Figure 3. However, the interviewer does **not** read the sources out loud for confirmation or negation. Instead, the interviewer places an x to the left of each source that the respondent mentions (possibly with the aid of bills and statements) and then enters the payment amount (if known) to the right of each source. The computer automatically checks to see if payments sum to the reported cost. However, the respondent is not pressed hard to reconcile any discrepancy.

It is important to note that there are two categories of payment data. The actual payment amounts carry the most information, but the x's on the left side of the screen also carry information. As an example, the beneficiary may know that Medicaid paid something toward the cost of the container but not know the amount paid by Medicaid. The algorithm was designed to preserve both types of partial data, as well as cost data.

3. Notation

Let $\delta=(\delta_1,..., \delta_S)$ where $\delta_i=1$ if the i-th source is known to have made a payment, $\delta_i=0$ if the i-th

[1] The authors are all employed at Westat, Inc., Rockville, MD. The work was supported by the Office of the Actuary in the Health Care Financing Administration under contract #500-90-0007.

component is known not to have made a payment. Given the structure of the interview, setting the delta's was not entirely straightforward. If there was an x

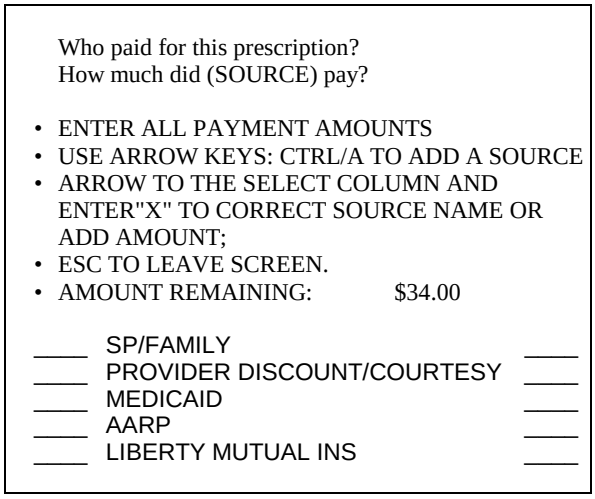


Figure 1. CAPI screen prior to entering payment data

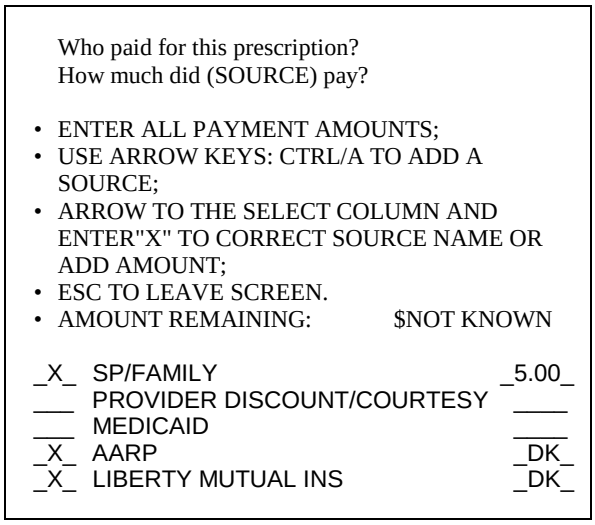


Figure 2. CAPI screen after entering partial payment data

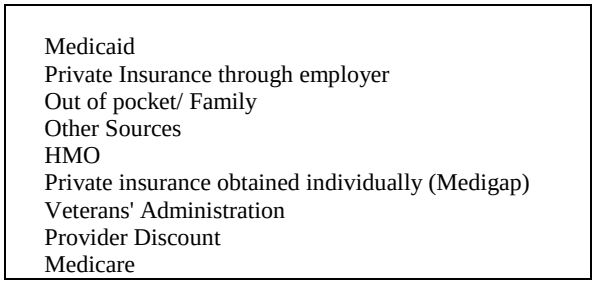


Figure 3. Sources of Payment

next to the source, then it was clear that the corresponding delta should be 1 (whether or not the payment amount was known). Also, if the insurance and program participation section of the questionnaire indicated that a person wasn't eligible for a particular source category, then it was clear that the corresponding delta should be 0. If, however, a person was eligible for coverage by source i, but there was no x next to source i, then determination of delta was more difficult. The rule we used was to set that delta component to 0 if the reported payment amounts summed to the cost or if analysis felt it unlikely that this source would pay given payments by other sources. Otherwise, that delta component was left missing. Let $h=(h_1,..., h_s)$ where $h_i=1$ if δ_i is "observed" and 0 otherwise.

Let $Y=(Y_1,..., Y_s)$ where Y_i is the payment by the i-th source. Let $g=(g_1,...,g_s)$ where $g_i=1$ if Y_i is observed and 0 otherwise. Let Y_+ be the total cost of the medicine container and g_+ indicate whether Y_+ is observed.

The total vector to be completed for each container of medicine is $\zeta=(\delta,Y,Y_+)$. Note that $h_i=0$ implies that $g_i=0$. Subject to that restriction, almost any pattern of missingness is possible.

To aid in the imputation, the analyst will typically have a set of background variables available which provide predictive information about the composition. In this application, the most important auxiliary data that we had for imputing δ was whether the person was eligible for assistance from each of the payment sources during the period when the purchase was made. We frequently also had information about the prescription such as name and strength, but these data were fully exploited in a separate exogenous imputation process that preceded our imputation work and is described below. In addition, we had a great wealth of background variables available at the person level such as income, education, region, metropolitan status, and so on. These person-level variables were thought to be important in imputing cost and payment amounts but unimportant in terms of predicting payment status (the delta vector) for each event. Without going into more detail about these background variables here, let X be a vector of background variables that are available for each event.

Let Ω_h be the set of distinct values of h realized in the sample. Let Ω_δ be the set of distinct values of δ realized in the sample.

The unique feature of compositional data that makes them so difficult to impute is that they must obey two constraints:

$0 \le Y_i \le Y_+$ for every i and (1)

$\Sigma_i Y_i = Y_+$. (2)

In this application where some information is contained in the delta vector, it is also necessary to have the constraints that

$\delta_i=0$ iff $Y_i=0$ for every i, and (3)
$Y_i>0$ implies $\delta_i=1$ for every i.

4. Data Editing and Exogenous Imputation

The raw data were not very amenable to imputation. A very intensive editing phase had to be carried out prior to imputation. Interviewers were encouraged to enter all relevant data about health care events that respondents shared with them. The data were collected over five interviews. The entire process of settling a large bill could take months and generate a lot of paperwork. As time elapsed since the health care event, it was not unusual for respondents to first share receipts with the interviewer, then insurance statements, then explanations of benefits from HCFA, then more insurance statements. Account statements from providers after insurance statements might also have been shown to the interviewer. Insurance companies might initially have rejected claims and then paid them upon appeal. Interviewers were trained to extract the best information from the paperwork submitted at a single interview, but there was less control over the entering of duplicate and/or contradictory data across interviews. Partly this was due to changes in interviewer assignments across time and partly it was due to a deliberate design decision to gather as many data as possible while in the beneficiaries' homes with the intent to sort it out later. An algorithm was developed by analysts at Westat to sift through the multiple reports of cost for the same event and to pull together the data that was felt to be best.

This was only half the editing battle, however. The other half involved cases where respondents submitted claims to insurance companies or other payment sources for multiple purchases of medicine (with or without other health care claims). Statements resulting from these claims often did not break the cost, copayment or deductible information down to the event level. The interviewer was trained to just enter the summary payment information for the claim as a whole. Staff at HCFA worked out a strategy to apportion the cost and payment information back to individual events. As part of this effort, they developed a means of exogenously imputing a reasonable total charge for many purchases based upon the name, strength, and volume of the purchase and industry data on average prices.[2] Thus, at the end of months of concerted effort by others, we received a database where there was exactly one record per container of medicine. On that record was the best payment information that could be salvaged from respondent reports and the price indicated by the respondent or a price exogenously imputed by HCFA. The only records for which cost was still missing were those for which the respondent was unable to recall the name. Since interviewers were trained to only enter data about prescription drugs, the assumption was made that these containers of "little yellow pills" and "heart pills" were truly prescription drugs and not over the counter medications.

5. Missing Data Rates after Editing and Exogenous Imputation

Table 1 shows the missing data rates on the delta vectors and for the actual payment amounts given that a source is known to have made a contribution. Examining the missing rates for payment status, we see that for the most part, respondents know who paid for their prescription medicine -- or rather, we can rule out payors on the basis of insurance and program participation data. The greatest uncertainty concerns whether the beneficiary had to make a payment out of pocket and whether there was a provider discount. This is strongly influenced by the way in which the data were collected and edited. If known payments didn't add to the total charge and if there was no mention of self payment or discount, then we generally assumed that these payment sources were possible and hence missing.[3] The pattern of uncertainty is quite different for payment amounts by known payors as is shown in the last column of Table 1. More than 75 percent of respondents could give us the amount of out-of-pocket payments and the amount of any discount. Knowledge about payments by other sources was generally weak. (The low nonresponse rate for Medicaid is a result of edit rules and the exogenous imputation of charges rather than of respondent knowledge.)

To place these item nonresponse rates in context, although the rates are high compared to those typically experienced on surveys on other subject matters (such as labor force behavior), we do not view them as extraordinarily high for a consumer expenditure survey. People have a difficult time saving all receipts and bills for us over the typical four-month span between interviews. The few dollars spent as a co-payment for one container of medicine three months earlier do not constitute a very salient

[2] Industry data on wholesale prices are available to HCFA for the administration of the Medicaid system. HCFA adjusted the wholesale prices to bring them up to likely retail levels with different factors depending upon the known payers. For example, it was assumed that Medicaid, HMOs, and VA usually paid considerably less for the same container of medicine than did individual beneficiaries at their local pharmacies.

[3] There were some exceptions to this general rule. If Medicaid was mentioned as a payer, then unmentioned sources were ruled out except HMO. Also provider discount was ruled out unless mentioned when the VA or an HMO was a known payer.

event in the typical respondent's memory. Furthermore, for those who are good about collecting receipts, many let them accumulate for months before submitting claims to insurance companies. Even with the longitudinal nature of the MCBS, it is difficult to track these claims over time. Most importantly, certain classes of beneficiaries have no knowledge of the cost of their prescription medicine; this is true for those who receive their drugs from the VA, from HMOs, through Medicaid, and through other public programs.

Table 1. Missing data rates

Payment source	Frequency of unknown payment status (Yes/No)[4] (%)	Frequency of unknown payment amount given payment status = Yes (%)
Medicaid	3.1	27.7
Private insurance provided by employer	5.2	67.1
Sample person and/or family (out of pocket)	11.5	23.6
Other sources	0.1	86.6
HMO	2.1	55.7
Private insurance individually purchased	2.1	62.0
Veterans' Administration	0.0	72.1
Provider discount	32.5	18.1
Medicare	0.0	78.5
Total charge	n/a	14.0

6. Patterns of Missingness in MCBS Prescription Medicine Data and the Decision to Impute

Despite the high missing data rates shown above, the majority of prescriptions were fully resolved after editing and exogenous imputation in the sense that payments agreed with charge. Furthermore, there were at least some data about every prescription in the sense that it was always possible to at least rule out one or more sources. Frequently, the data on the incomplete cases such as copayment amounts were useful and important.

A wide variety of approaches could have been adopted to deal with the incomplete cases. One approach would have been to discard the partial data (available on close to 50 percent of prescriptions) and then to either make up all the data about these prescriptions or to develop some sort of event-level weight that could be applied to complete records to weight up to the person level. Event-level weighting would have been problematic in that some people had no completely reported prescriptions at all. It would have been necessary to drop these people from analytic files altogether and give their weights to others. (In fact, a more extreme approach could have been taken of dropping everyone with at least one incomplete prescription, but that would have resulted in a very small analytic file. The exact number hasn't been tabulated yet, but it appears that the vast majority of people had at least one incomplete prescription.) Besides the confusion that event-level weights would have created among users, it was felt that the partial prescription reports often had valuable data within them that ought to be preserved.

Another approach would have been to discard just the partial payment data on the incomplete cases, keeping the total charge where it was known or exogenously imputed. This approach (similar to the one used for the 1987 National Medical Expenditure Survey) is very simple to implement since the cost can be imputed without any fear of contradicting the payment data (such as would be the case if a cost was imputed to be less than a payment). After imputing cost, the payment data can be imputed on a percentage basis using cases with complete payment patterns and similar insurance status as donors. This approach was considered and rejected out of the desire to preserve as much of the respondent-provided data as possible.

We wanted an approach that would preserve all the partial data (at least the partial data that were internally consistent), and build an internally consistent cost-payment report for each individual prescription while not distorting any important multivariate relationships as so often occurs with imputation.

Preserving the partial data while building an internally consistent record and not distorting distributions means conditioning upon important aspects of the partial data. This posed an enormous challenge since there were a total of 90 distinct patterns in the delta matrix prior to imputation for cases where the total charge was missing and 82 where the total charge was known The next section describes how this challenge was met.

7. The Skeleton of the Algorithm

[4] As discussed in the text, nonresponse on payment status is difficult to measure since the failure to mention a source can either reflect a definite nonpayment status for a source or a lack of knowledge. Edit rules were required to interpret the failure to mention as either a "no" or as a "don't know."

The algorithm has an iterative aspect that was inspired by Gibbs Sampling. However, it is not a strict application of that technique.

The first step is to make sure that the reported data obey the constraints and that nothing can be filled in by simple subtraction or addition. A variety of violations were found in the reported data. These violations were resolved in a separate editing step. The details of that editing will be covered in a forthcoming technical report.

The second step is to impute δ. This is done slightly differently depending upon whether the total cost is known and whether there are any known payors with unknown amounts. However, the basic idea is the same: For each element h of Ω_h, conduct a separate hot-deck run to impute the missing portion of δ, where the donors are chosen from among those cases that are already complete, the donors and missing cases are matched on X, the observed components of δ, and other available data. If the total cost is known, then that constitutes other available data that can be added to the match criteria (roughened into broad categories). If total cost is known and every known payor has a known amount, then the amount of money that must be covered by the missing deltas also constitutes other available data. Given the size of Ω_h and the three possibilities of reporting in Y and Y_+ for each element of Ω_h, a total number of 123 hot-decks were required for this step.[5]

The third step is to come up with an initial feasible solution for Y and Y_+ without worrying about how good the solution is. An initial solution is one where Y and Y_+ are complete, obey the constraints, and are consistent with δ. The hope is that, due to the iterative nature of the procedure, the starting solution is not very important. We used two different methods to complete ζ depending upon g. If $g_+=0$ (i.e., Y_+ is missing), then we sequentially imputed each corresponding Y_i with a simple hot-deck where δ_i and X were the conditioning variables. After completion of Y, we imputed Y_+ as the sum of the imputed and reported Y_i. If, on the other hand, $g_+=1$, then we counted up the number of missing Y_i thought to be positive as $m=\Sigma_i\delta_i(1-g_i)$ and set each of the positive missing $Y_i=(Y_+-Y_{R+})/m$, where $Y_{R+}=\Sigma_i\delta_j g_j Y_j$ is the sum of reported elements of Y.

The fourth step is to re-impute Y_1 for each case where Y_1 and Y_+ were both originally missing. This is done with a hot deck conditioned upon the sum of the other components of Y and on X. After Y_1 is re-imputed, its new value is added on to the sum of the other components to obtain a new value for Y_+. This step is repeated for each of the Y_i. The motivation for the step is to improve the pair-wise consistency of the individual Y_i with the total, Y_+.

The fifth step is to re-impute the division of Y_1+Y_2 between Y_1 and Y_2 for all cases where both Y_1 and Y_2 were originally missing. This is done with a hot deck conditioned on Y_1+Y_2 and X. The hot deck actually imputes $P_1=Y_1/(Y_1+Y_2)$. The program then computes appropriate new values of Y_1 and Y_2. This step is repeated for each possible pair of components of Y. The motivation for the step is to improve the pair-wise consistency of the components of Y.

The fourth and fifth steps are then iterated until the national total number of dollars paid by each source stabilizes. The word "stabilizes" was chosen here rather than "converges," because it is not clear how to even define convergence in this setting. On each iteration, payments and charges are being resampled from similar cases. Since within each pool of similar donors, there is some variation, the individual values and, to a lesser extent, the national means will continue to fluctuate indefinitely.

8. Results

The algorithm was stopped after five iterations. Table 2 shows some summary information about CPU times and measures of change across iterations. The CPU times were much more modest than expected but still significant. The change statistics indicate that changes at the national level on broad measures were fairly small by the fifth iteration. This is comforting but doesn't exclude significant instability for more narrow measures. For example, the average Medicare payment changed by 5 percent from iteration 4 to iteration 5. This was perhaps not too surprising given that Medicare pays for only 1 or 2 prescriptions from every thousand and that the payment can be large when it does pay, but it does leave open the question of convergence in some broad sense.

[5] The maximum possible number of runs is 3.2s, or 1536 in this application with s=9. If s had been larger, this procedure may not have been practical. Judkins, Hubbell, and England (1993) discuss some possible alternatives.

Table 2. Selected results of applying algorithm to prescription medicine data

	CPU hours on IBM mainframe	Relative change in average cost per container (%)	Percentage of national dollars shifted among sources
Initial Solution	2.8	n/a	n/a
Iteration 1	0.9	-1.43	17.15
Iteration 2	0.9	0.21	0.58
Iteration 3	0.9	-0.09	0.39
Iteration 4	0.9	0.04	0.39
Iteration 5	1.1	0.05	0.24
Total	7.5	n/a	n/a

The covariance matrix of the delta vector, the covariance matrix of the Y vector, and the average payment amounts for each delta pattern were monitored as well throughout the imputation process. We noted that some correlations did change. It is difficult to know whether these changes were good or bad, but we can say that there was very little attenuation of corrrelations between payment amounts by different sources. Those that were negative tended to stay negative and those that were positive tended to stay positive. In fact, some correlations increased in strength as a result of the imputation. In particular, the correlation between the payment amount by private employer-provided insurance and the total charge was noticeably stronger after imputation. We hope to be able to share these more detailed results in a full technical report at a later date.

9. Limitations

Two limitations of the algorithm were noted. The first concerns instances where the observed data set does not contain any completely observed relevant data. The second concerns estimation of precision on the fully imputed dataset.

The algorithm was designed to preserve partial data by building a consistent financial reckoning around reported data. Furthermore, it was designed to do this in a way that minimally distorts observed payment patterns and relationships between amounts paid by various sources. To accomplish this, it relied upon observed distributions on similar but fully reported cases to decide how to identify payors and allocate dollars across sources. When there were no similar cases that were fully observed, the algorithm created some very unintuitive results. Only one example of this has been detected so far, but there are probably others waiting to be discovered. The example involved Medicaid payments for insulin. There was not a single Medicaid respondent who could tell us either the cost or the Medicaid payment for insulin. The hot-deck program that was used to implement the program has an automatic feature for dealing with cells that have no donors. It borrows from the cell that is closest to the deficient cell in terms of hierarchical agreement on the background variables. In this case, the nearest cell was not an appropriate source of donors. As a result of this, the insulin data were redone separately from the true prescription drug data. The weakness in the algorithm that we have discovered thus concerns situations where no similar person in the sample could provide any useful data. In such situations, external knowledge must be brought into the imputation process.

Turning attention to the second limitation, users of the fully imputed dataset may be lulled into a false sense of security. A large percentage of total dollars and their allocation across payors is imputed. Yet, the user will appear to have complete data on close to 250,000 containers of prescriptions medicine for about 10,000 Medicare beneficiaries. Standard errors estimated from this dataset by conventional means will not be very accurate. We have provided resampling weights so that the variance estimates can be inflated for the complex sample design, but we have no very satisfactory way of adjusting estimated standard errors for the imputation process. Clearly, estimated standard errors will tend to be much too small. A burgeoning literature exists on methods for fully reflecting uncertainty in imputed datasets, but none of these methods seemed developed enough to use in conjunction with this new approach to imputing compositional data. For the moment, the best we can advise users is to inflate estimated variances by the inverse of the observed item response rate. A related question is what sort of variance to associate with the exogenous imputation process that was carried out.

10. Conclusions

The algorithm succeeded in creating a full set of internally consistent cost and payment records while discarding very little partial data. Indeed, the only partial data that were discarded were those that were already internally inconsistent prior to imputation. Some distributional changes were observed, but if that was not the case, then there would have been little point in doing the imputation. In other words, if analysis of the fully imputed dataset yielded the same results as analysis of just the fully reported cases, then the only reason to do the imputation would be to make tabulations easier for

analysts. Computer requirements were intensive but not as intensive as feared. We plan to continue to use the algorithm to impute cost and payment data for other medical services.

References

Hahn, B. and Lefkowitz, D. (1992). *Annual expenses and sources of payment for health care services* (AHCPR Pub. No. 93-0007). National Medical Expenditure Survey Research Findings 14, Agency for Health Care Policy and Research, Rockville, MD: Public Health Service.

Judkins, D., Hubbell, K.A. and England, A.M. (1993). "The Imputation of Compositional Data." *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 458-462.

IMPUTATION OF MEDICAL COST AND PAYMENT DATA

Amy M. England[1], Katie A. Hubbell[1], David R. Judkins[1], Svetlana Ryaboy[1]
Westat, Inc., 1650 Research Blvd., Rockville, MD 20850

Keywords: Gibbs Sampling, Hot Deck Imputation, Compositional Data[1]

Medical cost and payment data are the primary focus of the Medicare Current Beneficiary Survey (MCBS). These data are compositional data (data where a finite series of random variables are non-negative and sum to another random variable). There is a large variety of missing patterns that are neither nested nor ignorable. A paper from last year presented a new technique for creating a complete set of compositional data while preserving all partial data and maintaining many types of consistency. This year, we present the results of applying the method to actual MCBS data on prescription drugs. Since the method is known to be extremely CPU intensive, a primary point of interest will be the feasibility of applying the method to a dataset with about 245,000 records and nine possible payment sources.

1. Introduction

The imputation of costs and payment sources for prescription medicines is a critical area for the Medicare Current Beneficiary Survey (MCBS) given the ongoing national debate about whether to expand Medicare coverage to include prescription medicines. There were a substantial number of partially complete reports about purchases of containers of prescription medicine. One solution is to impute the cost where necessary, discard partial payment data, and impute whole payment vectors as proportions to be applied to the cost. This solution was used for example on the 1987 National Medical Expenditure Survey (Hahn and Lefkowitz, 1992, p22). Judkins, Hubbell and England (1993), presented an alternate solution that allows the retention of all partial data payment and cost data. They presented an evaluation of the algorithm on an artificial example. That evaluation focused on the ability of the algorithm to minimize nonresponse bias. In this paper, we evaluate the algorithm in terms of practicality by presenting the results of its application to the 245,000 records for individual containers of prescription medicine in the 1992 MCBS.

In the following sections, we review briefly how prescription drug data are collected in the MCBS, define some notation, present some information on the patterns of missingness observed in MCBS prescription data, review the algorithm (some improvements have been made over the version presented last year), and, finally, present results and ideas for future improvements.

2. Data Collection

The MCBS has a modified panel design where a core panel is supplemented once a year with new additions to the eligible universe and additional beneficiaries from the original cohort so as to maintain cross-sectional precision despite deaths and attrition in the panel. Interviews are conducted roughly every four months. The reference period for each interview extends from the date of the prior interview to the date of current interview. Data are collected about the utilization of health care services, the costs of these services, and expenditures (personal and third-party) for these services.

MCBS data are collected by CAPI (computer assisted personal interview). Interviewers carry laptop computers into the homes of Medicare beneficiaries and run a program that guides them through the interview. Figure 1 mimics a typical screen for collecting information about payments for a health care event after the cost has been determined. Figure 2 shows how it might look after completion. Note that the program presents a list of possible payment sources for the event and that the list is tailored to the beneficiary's insurance status and program participation. The payment sources mentioned by respondents were grouped into the nine categories shown in Figure 3. However, the interviewer does **not** read the sources out loud for confirmation or negation. Instead, the interviewer places an x to the left of each source that the respondent mentions (possibly with the aid of bills and statements) and then enters the payment amount (if known) to the right of each source. The computer automatically checks to see if payments sum to the reported cost. However, the respondent is not pressed hard to reconcile any discrepancy.

It is important to note that there are two categories of payment data. The actual payment amounts carry the most information, but the x's on the left side of the screen also carry information. As an example, the beneficiary may know that Medicaid paid something toward the cost of the container but not know the amount paid by Medicaid. The algorithm was designed to preserve both types of partial data, as well as cost data.

3. Notation

Let $\delta=(\delta_1,..., \delta_S)$ where $\delta_i=1$ if the i-th source is known to have made a payment, $\delta_i=0$ if the i-th

[1] The authors are all employed at Westat, Inc., Rockville, MD. The work was supported by the Office of the Actuary in the Health Care Financing Administration under contract #500-90-0007.

component is known not to have made a payment. Given the structure of the interview, setting the delta's was not entirely straightforward. If there was an x

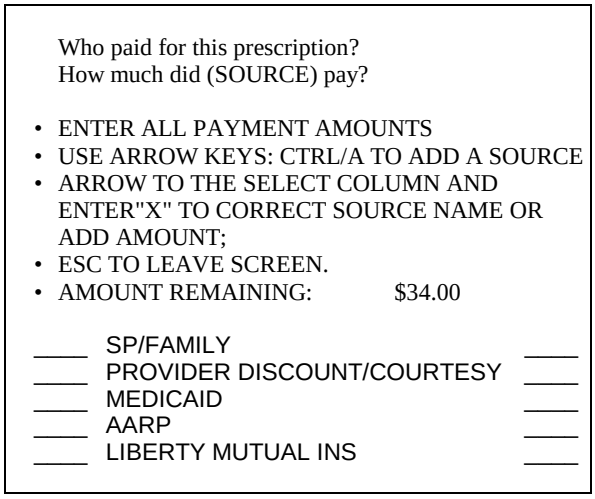


Figure 1. CAPI screen prior to entering payment data

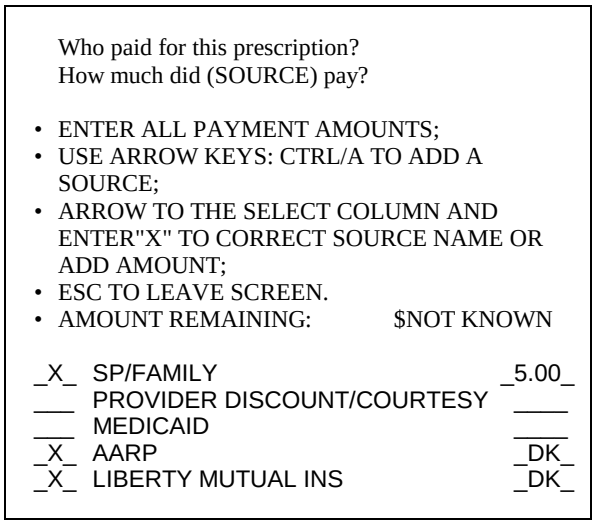


Figure 2. CAPI screen after entering partial payment data

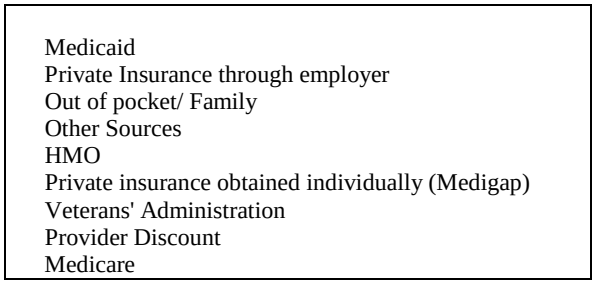


Figure 3. Sources of Payment

next to the source, then it was clear that the corresponding delta should be 1 (whether or not the payment amount was known). Also, if the insurance and program participation section of the questionnaire indicated that a person wasn't eligible for a particular source category, then it was clear that the corresponding delta should be 0. If, however, a person was eligible for coverage by source i, but there was no x next to source i, then determination of delta was more difficult. The rule we used was to set that delta component to 0 if the reported payment amounts summed to the cost or if analysis felt it unlikely that this source would pay given payments by other sources. Otherwise, that delta component was left missing. Let $h=(h_1,..., h_s)$ where $h_i=1$ if δ_i is "observed" and 0 otherwise.

Let $Y=(Y_1,..., Y_s)$ where Y_i is the payment by the i-th source. Let $g=(g_1,...,g_s)$ where $g_i=1$ if Y_i is observed and 0 otherwise. Let Y_+ be the total cost of the medicine container and g_+ indicate whether Y_+ is observed.

The total vector to be completed for each container of medicine is $\zeta=(\delta,Y,Y_+)$. Note that $h_i=0$ implies that $g_i=0$. Subject to that restriction, almost any pattern of missingness is possible.

To aid in the imputation, the analyst will typically have a set of background variables available which provide predictive information about the composition. In this application, the most important auxiliary data that we had for imputing δ was whether the person was eligible for assistance from each of the payment sources during the period when the purchase was made. We frequently also had information about the prescription such as name and strength, but these data were fully exploited in a separate exogenous imputation process that preceded our imputation work and is described below. In addition, we had a great wealth of background variables available at the person level such as income, education, region, metropolitan status, and so on. These person-level variables were thought to be important in imputing cost and payment amounts but unimportant in terms of predicting payment status (the delta vector) for each event. Without going into more detail about these background variables here, let X be a vector of background variables that are available for each event.

Let Ω_h be the set of distinct values of h realized in the sample. Let Ω_δ be the set of distinct values of δ realized in the sample.

The unique feature of compositional data that makes them so difficult to impute is that they must obey two constraints:

$0 \le Y_i \le Y_+$ for every i and (1)

$\Sigma_i Y_i = Y_+$. (2)

In this application where some information is contained in the delta vector, it is also necessary to have the constraints that

$\delta_i=0$ iff $Y_i=0$ for every i, and (3)
$Y_i>0$ implies $\delta_i=1$ for every i.

4. Data Editing and Exogenous Imputation

The raw data were not very amenable to imputation. A very intensive editing phase had to be carried out prior to imputation. Interviewers were encouraged to enter all relevant data about health care events that respondents shared with them. The data were collected over five interviews. The entire process of settling a large bill could take months and generate a lot of paperwork. As time elapsed since the health care event, it was not unusual for respondents to first share receipts with the interviewer, then insurance statements, then explanations of benefits from HCFA, then more insurance statements. Account statements from providers after insurance statements might also have been shown to the interviewer. Insurance companies might initially have rejected claims and then paid them upon appeal. Interviewers were trained to extract the best information from the paperwork submitted at a single interview, but there was less control over the entering of duplicate and/or contradictory data across interviews. Partly this was due to changes in interviewer assignments across time and partly it was due to a deliberate design decision to gather as many data as possible while in the beneficiaries' homes with the intent to sort it out later. An algorithm was developed by analysts at Westat to sift through the multiple reports of cost for the same event and to pull together the data that was felt to be best.

This was only half the editing battle, however. The other half involved cases where respondents submitted claims to insurance companies or other payment sources for multiple purchases of medicine (with or without other health care claims). Statements resulting from these claims often did not break the cost, copayment or deductible information down to the event level. The interviewer was trained to just enter the summary payment information for the claim as a whole. Staff at HCFA worked out a strategy to apportion the cost and payment information back to individual events. As part of this effort, they developed a means of exogenously imputing a reasonable total charge for many purchases based upon the name, strength, and volume of the purchase and industry data on average prices.[2] Thus, at the end of months of concerted effort by others, we received a database where there was exactly one record per container of medicine. On that record was the best payment information that could be salvaged from respondent reports and the price indicated by the respondent or a price exogenously imputed by HCFA. The only records for which cost was still missing were those for which the respondent was unable to recall the name. Since interviewers were trained to only enter data about prescription drugs, the assumption was made that these containers of "little yellow pills" and "heart pills" were truly prescription drugs and not over the counter medications.

5. Missing Data Rates after Editing and Exogenous Imputation

Table 1 shows the missing data rates on the delta vectors and for the actual payment amounts given that a source is known to have made a contribution. Examining the missing rates for payment status, we see that for the most part, respondents know who paid for their prescription medicine -- or rather, we can rule out payors on the basis of insurance and program participation data. The greatest uncertainty concerns whether the beneficiary had to make a payment out of pocket and whether there was a provider discount. This is strongly influenced by the way in which the data were collected and edited. If known payments didn't add to the total charge and if there was no mention of self payment or discount, then we generally assumed that these payment sources were possible and hence missing.[3] The pattern of uncertainty is quite different for payment amounts by known payors as is shown in the last column of Table 1. More than 75 percent of respondents could give us the amount of out-of-pocket payments and the amount of any discount. Knowledge about payments by other sources was generally weak. (The low nonresponse rate for Medicaid is a result of edit rules and the exogenous imputation of charges rather than of respondent knowledge.)

To place these item nonresponse rates in context, although the rates are high compared to those typically experienced on surveys on other subject matters (such as labor force behavior), we do not view them as extraordinarily high for a consumer expenditure survey. People have a difficult time saving all receipts and bills for us over the typical four-month span between interviews. The few dollars spent as a co-payment for one container of medicine three months earlier do not constitute a very salient

[2] Industry data on wholesale prices are available to HCFA for the administration of the Medicaid system. HCFA adjusted the wholesale prices to bring them up to likely retail levels with different factors depending upon the known payers. For example, it was assumed that Medicaid, HMOs, and VA usually paid considerably less for the same container of medicine than did individual beneficiaries at their local pharmacies.

[3] There were some exceptions to this general rule. If Medicaid was mentioned as a payer, then unmentioned sources were ruled out except HMO. Also provider discount was ruled out unless mentioned when the VA or an HMO was a known payer.

event in the typical respondent's memory. Furthermore, for those who are good about collecting receipts, many let them accumulate for months before submitting claims to insurance companies. Even with the longitudinal nature of the MCBS, it is difficult to track these claims over time. Most importantly, certain classes of beneficiaries have no knowledge of the cost of their prescription medicine; this is true for those who receive their drugs from the VA, from HMOs, through Medicaid, and through other public programs.

Table 1. Missing data rates

Payment source	Frequency of unknown payment status (Yes/No)[4] (%)	Frequency of unknown payment amount given payment status = Yes (%)
Medicaid	3.1	27.7
Private insurance provided by employer	5.2	67.1
Sample person and/or family (out of pocket)	11.5	23.6
Other sources	0.1	86.6
HMO	2.1	55.7
Private insurance individually purchased	2.1	62.0
Veterans' Administration	0.0	72.1
Provider discount	32.5	18.1
Medicare	0.0	78.5
Total charge	n/a	14.0

6. Patterns of Missingness in MCBS Prescription Medicine Data and the Decision to Impute

Despite the high missing data rates shown above, the majority of prescriptions were fully resolved after editing and exogenous imputation in the sense that payments agreed with charge. Furthermore, there were at least some data about every prescription in the sense that it was always possible to at least rule out one or more sources. Frequently, the data on the incomplete cases such as copayment amounts were useful and important.

A wide variety of approaches could have been adopted to deal with the incomplete cases. One approach would have been to discard the partial data (available on close to 50 percent of prescriptions) and then to either make up all the data about these prescriptions or to develop some sort of event-level weight that could be applied to complete records to weight up to the person level. Event-level weighting would have been problematic in that some people had no completely reported prescriptions at all. It would have been necessary to drop these people from analytic files altogether and give their weights to others. (In fact, a more extreme approach could have been taken of dropping everyone with at least one incomplete prescription, but that would have resulted in a very small analytic file. The exact number hasn't been tabulated yet, but it appears that the vast majority of people had at least one incomplete prescription.) Besides the confusion that event-level weights would have created among users, it was felt that the partial prescription reports often had valuable data within them that ought to be preserved.

Another approach would have been to discard just the partial payment data on the incomplete cases, keeping the total charge where it was known or exogenously imputed. This approach (similar to the one used for the 1987 National Medical Expenditure Survey) is very simple to implement since the cost can be imputed without any fear of contradicting the payment data (such as would be the case if a cost was imputed to be less than a payment). After imputing cost, the payment data can be imputed on a percentage basis using cases with complete payment patterns and similar insurance status as donors. This approach was considered and rejected out of the desire to preserve as much of the respondent-provided data as possible.

We wanted an approach that would preserve all the partial data (at least the partial data that were internally consistent), and build an internally consistent cost-payment report for each individual prescription while not distorting any important multivariate relationships as so often occurs with imputation.

Preserving the partial data while building an internally consistent record and not distorting distributions means conditioning upon important aspects of the partial data. This posed an enormous challenge since there were a total of 90 distinct patterns in the delta matrix prior to imputation for cases where the total charge was missing and 82 where the total charge was known The next section describes how this challenge was met.

7. The Skeleton of the Algorithm

[4] As discussed in the text, nonresponse on payment status is difficult to measure since the failure to mention a source can either reflect a definite nonpayment status for a source or a lack of knowledge. Edit rules were required to interpret the failure to mention as either a "no" or as a "don't know."

The algorithm has an iterative aspect that was inspired by Gibbs Sampling. However, it is not a strict application of that technique.

The first step is to make sure that the reported data obey the constraints and that nothing can be filled in by simple subtraction or addition. A variety of violations were found in the reported data. These violations were resolved in a separate editing step. The details of that editing will be covered in a forthcoming technical report.

The second step is to impute δ. This is done slightly differently depending upon whether the total cost is known and whether there are any known payors with unknown amounts. However, the basic idea is the same: For each element h of Ω_h, conduct a separate hot-deck run to impute the missing portion of δ, where the donors are chosen from among those cases that are already complete, the donors and missing cases are matched on X, the observed components of δ, and other available data. If the total cost is known, then that constitutes other available data that can be added to the match criteria (roughened into broad categories). If total cost is known and every known payor has a known amount, then the amount of money that must be covered by the missing deltas also constitutes other available data. Given the size of Ω_h and the three possibilities of reporting in Y and Y_+ for each element of Ω_h, a total number of 123 hot-decks were required for this step.[5]

The third step is to come up with an initial feasible solution for Y and Y_+ without worrying about how good the solution is. An initial solution is one where Y and Y_+ are complete, obey the constraints, and are consistent with δ. The hope is that, due to the iterative nature of the procedure, the starting solution is not very important. We used two different methods to complete ζ depending upon g. If $g_+=0$ (i.e., Y_+ is missing), then we sequentially imputed each corresponding Y_i with a simple hot-deck where δ_i and X were the conditioning variables. After completion of Y, we imputed Y_+ as the sum of the imputed and reported Y_i. If, on the other hand, $g_+=1$, then we counted up the number of missing Y_i thought to be positive as $m=\Sigma_i\delta_i(1-g_i)$ and set each of the positive missing $Y_i=(Y_+-Y_{R+})/m$, where $Y_{R+}=\Sigma_i\delta_j g_j Y_j$ is the sum of reported elements of Y.

The fourth step is to re-impute Y_1 for each case where Y_1 and Y_+ were both originally missing. This is done with a hot deck conditioned upon the sum of the other components of Y and on X. After Y_1 is re-imputed, its new value is added on to the sum of the other components to obtain a new value for Y_+. This step is repeated for each of the Y_i. The motivation for the step is to improve the pair-wise consistency of the individual Y_i with the total, Y_+.

The fifth step is to re-impute the division of Y_1+Y_2 between Y_1 and Y_2 for all cases where both Y_1 and Y_2 were originally missing. This is done with a hot deck conditioned on Y_1+Y_2 and X. The hot deck actually imputes $P_1=Y_1/(Y_1+Y_2)$. The program then computes appropriate new values of Y_1 and Y_2. This step is repeated for each possible pair of components of Y. The motivation for the step is to improve the pair-wise consistency of the components of Y.

The fourth and fifth steps are then iterated until the national total number of dollars paid by each source stabilizes. The word "stabilizes" was chosen here rather than "converges," because it is not clear how to even define convergence in this setting. On each iteration, payments and charges are being resampled from similar cases. Since within each pool of similar donors, there is some variation, the individual values and, to a lesser extent, the national means will continue to fluctuate indefinitely.

8. Results

The algorithm was stopped after five iterations. Table 2 shows some summary information about CPU times and measures of change across iterations. The CPU times were much more modest than expected but still significant. The change statistics indicate that changes at the national level on broad measures were fairly small by the fifth iteration. This is comforting but doesn't exclude significant instability for more narrow measures. For example, the average Medicare payment changed by 5 percent from iteration 4 to iteration 5. This was perhaps not too surprising given that Medicare pays for only 1 or 2 prescriptions from every thousand and that the payment can be large when it does pay, but it does leave open the question of convergence in some broad sense.

[5] The maximum possible number of runs is 3.2s, or 1536 in this application with s=9. If s had been larger, this procedure may not have been practical. Judkins, Hubbell, and England (1993) discuss some possible alternatives.

Table 2. Selected results of applying algorithm to prescription medicine data

	CPU hours on IBM mainframe	Relative change in average cost per container (%)	Percentage of national dollars shifted among sources
Initial Solution	2.8	n/a	n/a
Iteration 1	0.9	-1.43	17.15
Iteration 2	0.9	0.21	0.58
Iteration 3	0.9	-0.09	0.39
Iteration 4	0.9	0.04	0.39
Iteration 5	1.1	0.05	0.24
Total	7.5	n/a	n/a

The covariance matrix of the delta vector, the covariance matrix of the Y vector, and the average payment amounts for each delta pattern were monitored as well throughout the imputation process. We noted that some correlations did change. It is difficult to know whether these changes were good or bad, but we can say that there was very little attenuation of corrrelations between payment amounts by different sources. Those that were negative tended to stay negative and those that were positive tended to stay positive. In fact, some correlations increased in strength as a result of the imputation. In particular, the correlation between the payment amount by private employer-provided insurance and the total charge was noticeably stronger after imputation. We hope to be able to share these more detailed results in a full technical report at a later date.

9. Limitations

Two limitations of the algorithm were noted. The first concerns instances where the observed data set does not contain any completely observed relevant data. The second concerns estimation of precision on the fully imputed dataset.

The algorithm was designed to preserve partial data by building a consistent financial reckoning around reported data. Furthermore, it was designed to do this in a way that minimally distorts observed payment patterns and relationships between amounts paid by various sources. To accomplish this, it relied upon observed distributions on similar but fully reported cases to decide how to identify payors and allocate dollars across sources. When there were no similar cases that were fully observed, the algorithm created some very unintuitive results. Only one example of this has been detected so far, but there are probably others waiting to be discovered. The example involved Medicaid payments for insulin. There was not a single Medicaid respondent who could tell us either the cost or the Medicaid payment for insulin. The hot-deck program that was used to implement the program has an automatic feature for dealing with cells that have no donors. It borrows from the cell that is closest to the deficient cell in terms of hierarchical agreement on the background variables. In this case, the nearest cell was not an appropriate source of donors. As a result of this, the insulin data were redone separately from the true prescription drug data. The weakness in the algorithm that we have discovered thus concerns situations where no similar person in the sample could provide any useful data. In such situations, external knowledge must be brought into the imputation process.

Turning attention to the second limitation, users of the fully imputed dataset may be lulled into a false sense of security. A large percentage of total dollars and their allocation across payors is imputed. Yet, the user will appear to have complete data on close to 250,000 containers of prescriptions medicine for about 10,000 Medicare beneficiaries. Standard errors estimated from this dataset by conventional means will not be very accurate. We have provided resampling weights so that the variance estimates can be inflated for the complex sample design, but we have no very satisfactory way of adjusting estimated standard errors for the imputation process. Clearly, estimated standard errors will tend to be much too small. A burgeoning literature exists on methods for fully reflecting uncertainty in imputed datasets, but none of these methods seemed developed enough to use in conjunction with this new approach to imputing compositional data. For the moment, the best we can advise users is to inflate estimated variances by the inverse of the observed item response rate. A related question is what sort of variance to associate with the exogenous imputation process that was carried out.

10. Conclusions

The algorithm succeeded in creating a full set of internally consistent cost and payment records while discarding very little partial data. Indeed, the only partial data that were discarded were those that were already internally inconsistent prior to imputation. Some distributional changes were observed, but if that was not the case, then there would have been little point in doing the imputation. In other words, if analysis of the fully imputed dataset yielded the same results as analysis of just the fully reported cases, then the only reason to do the imputation would be to make tabulations easier for

analysts. Computer requirements were intensive but not as intensive as feared. We plan to continue to use the algorithm to impute cost and payment data for other medical services.

References

Hahn, B. and Lefkowitz, D. (1992). *Annual expenses and sources of payment for health care services* (AHCPR Pub. No. 93-0007). National Medical Expenditure Survey Research Findings 14, Agency for Health Care Policy and Research, Rockville, MD: Public Health Service.

Judkins, D., Hubbell, K.A. and England, A.M. (1993). "The Imputation of Compositional Data." *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 458-462.

Computer Matching of Medicare Current Beneficiary Survey Data With Medicare Claims

Franklin J. Eppig Jr. and Brad Edwards

The Medicare Current Beneficiary Survey (MCBS) is a continuous panel survey of Medicare beneficiaries in the United States.[1] Interviews are conducted three times a year with a sample of about 12,000 to collect information about the use and cost of health care services. All household interviews are conducted in person by computer-assisted personal interviewing (CAPI). In addition to the usual features of computer-assisted interviewing (CAI), the MCBS CAPI design includes extensive abstracting of documents, especially explanations of Medicare benefits and statements that reflect private insurance coverage for specific events. Because a critical MCBS goal is to estimate payments by various sources for services that Medicare covers (but does not pay in full), for each reported service, the survey attempts to identify the total charge (or the Medicare-approved charge, for participating providers) and the Medicare payment in order to determine the amount for which the Medicare beneficiary or other payment sources are responsible.

When a Medicare enrollee receives a Medicare-covered service, the medical provider submits a claim for payment directly to Medicare.[2] Even if the provider refuses to accept assignment and requires the patient to pay for the service and seek reimbursement from Medicare, the provider is still required to submit the claim for payment. After Medicare claims are processed for payment by Medicare's fiscal agents, they are forwarded to the National Claims History database (NCH). An estimated 97% of the Medicare claims are posted to NCH within a year, even with processing delays related to adjudication of disputed claims. Thus, NCH data provide a nearly complete picture of the Medicare utilization and reimbursements for all but the 6% of the Medicare population enrolled in capitate plans.

However, the NCH database contains no information about other payment sources for events covered by Medicare, nor does it include items/events that Medicare does not cover (such as most prescribed medicines or physician services for persons covered by Part A but not by Part B). The survey interviewer asks the beneficiary about all events and attempts to collect data on all payment sources and amounts for those events. The best estimate for total expenditures for all events is derived from a combination of the two data sources.

Objectives for Matching MCBS Survey Data and Medicare Claims

Matching survey data with claims data has two primary objectives: to adjust for underreporting of the use of health care services by survey respondents and to fill gaps and make corrections in the survey expenditure data.

Underreporting health care events has been a subject of considerable interest in the survey literature. Memory of specific events is prone to decay, and even the best efforts to probe respondents' memories and to assist their recall are unlikely to boost reporting to desirable levels, particularly for events that are not very salient and for recall periods that are very long. A person level comparison of survey-reported events with events in the Medicare claims can identify events that the respondent may have forgotten. Other events may be difficult or impossible for the respondent to report, not because of memory limitations, but because of the way the events are experienced. For instance, laboratory services may be classified as events in their own right, but the respondent may never be conscious of them—it's a mystery to the patient what happens to the blood once it's drawn. The Medicare records system, however, treats laboratory services like other events and services, so it is a better source for these "hidden" event categories.

Franklin J. Eppig Jr. is with the Health Care Financing Administration in Baltimore, Maryland. Brad Edwards is with Westat, Inc., in Rockville, Maryland.

[1]The Medicare program is a federal health insurance program for people 65 or older and certain disabled people. Approximately 34,000,000 Americans are enrolled in Medicare. Medicare Hospital Insurance, Part A, covers inpatient hospital care, inpatient care in a skilled nursing facility following a hospital stay, home health care, and hospice care. Medicare Medical Insurance, Part B, helps pay for doctors' services, outpatient hospital care, diagnostic test, durable medical equipment, ambulance services, and many other health services and supplies.

[2]This is not true for Medicare beneficiaries who are enrolled in capitate plans. Because their services are not provided on a fee-for-service basis no claim is submitted to Medicare for payment. As a result, Medicare administrative claims databases do not capture utilization and expenditures for medical services provided through capitate arrangements. In 1992 about 6% of those in the Medicare population were members of capitate plans.

Survey respondents experience even more difficulty in reporting expenditures for medical care than they do in reporting the occurrence of health care events. This is not surprising, especially given the complexity of the current health care financing systems in the United States. The survey respondent may be the best source for information on out-of-pocket payments, but the Medicare program is likely to be the best source for information on Medicare payments. For some events, such as inpatient hospital stays, Medicare and the provider may be the only sources for expenditure data because Medicare payments (under the Diagnostic Related Group [DRG] system) are not related to charges. Matching the survey events with the Medicare claims also allows us to check the respondent's reported expenditure data and to fill gaps when the respondent does not know the charges or the payment sources or amounts for covered services.

MCBS Matching Strategy

The first step in matching survey-reported events to Medicare claims is the association of all Medicare claims with a given sampled person. The MCBS design accommodates person level accumulation of Medicare claims data through its use of the Medicare health insurance claim number (HICN). The HICN appears on every Medicare claim submitted for payment and is the key to collecting all of a sampled person's Medicare claims. Since the MCBS sample is drawn from the Enrollment Data Base, the HICN for each sampled person is known prior to the start of field operations.

MCBS interviewers verify the sampled person's HICN during the initial interview using the HICN from the Enrollment Data Base. This circumvents the problems of misreporting and incorrect transcription associated with the collection of the HICN in the field. Having the correct HICN for each sampled person means that a sampled person's Medicare claims can be extracted from the NCH with complete accuracy.

A potential problem with using the HICN to capture an individual's Medicare claims is that a Medicare enrollee's HICN can change. For example, if an individual is entitled to Medicare benefits under both his or her own and a spouse's health insurance account, the HICN may change with the death of the spouse. The MCBS staff track claim number changes using internal Health Care Financing Administration (HCFA) files. This allows MCBS staff to capture all of an individual's Medicare claims regardless of claim number changes.

The next step is to determine the extent of overlap between the survey-reported events and claims data, which requires event level matching of survey data and claims data. Matching survey-reported data to Medicare claims at the event level is significantly more difficult than person level matching. Unlike the HICN at the person level, no data element or combination of data elements provides a consistent and reliably reported basis for conducting event level matches. Discrepancies in the reporting of the same event can occur because of differences in the perspective of the parties or the faulty recollection of specific details of events by respondents. The MCBS relies on Medicare explanation of benefits forms, insurance statements, and other receipts to assist the respondent's memory whenever possible (and as a source of other data elements, such as the claim control number, that were never stored in respondent memory). Often, however, the unaided memory of the respondent is the only source available for event details.

There are several other reasons for the lack of a consistent set of data for event matching. First, the MCBS does not capture a consistent set of variables for the different types of service. For example, the MCBS does not collect total charges or reimbursements for inpatient hospital events, since Medicare beneficiaries usually don't know this information. However, event total charges is a key match field for other survey event categories. Similarly, the MCBS does not capture date-of-service information for prescription drugs, home health events, and "other" medical expenses, but the date of service is a key match field for all other types of service. Second, there are different file layouts and different data elements on the Medicare claims for different service types. Third, for certain classes of beneficiaries (e.g., end stage renal disease [ESRD]) and certain repeat service situations, Medicare claims contain aggregate monthly billing information instead of event level data.

Differences in the categorization of medical services between the Medicare claims and the survey further complicate event level matching. The Medicare claims are essentially organized by type of provider, whereas the type of service categories used in the MCBS are more closely related to the way in which individuals think about the medical care they receive (see Figure 1). In matching the survey event to the Medicare claims data, MCBS staff

Figure 1. Comparison of Medicare claims categories with MCBS event categories

Medicare claims categories	MCBS event categories
Inpatient hospital	DU — Dental
Skilled nursing facility	ER — Emergency room
Hospice	IP — Inpatient hospital
services	
Home health agency	OP — Outpatient hospital services
Outpatient hospital	MP — Medical provider services
Part B physician/supplier	PM — Prescribed medicine
	HF — Home health services—friend
	HP — Home health services—prof.
	OM — Other medical
	IU — Institutional utilization
	SD — Separately billing doctors
	SL — Separately billing labs

frequently must match a Medicare claim category with multiple MCBS event categories and vice versa.

There are only 6 claims categories versus 12 MCBS event categories. Some of these discrepancies are readily explained. For example, dental services are not included in the claims list because Medicare does not cover most dental services. One of the most noteworthy categories missing from the claims list is emergency room services. In the Medicare claims system, emergency room services that are immediately followed by an inpatient stay are included in the DRG for the inpatient stay and thus are not associated with any separate charges or claims. Emergency room visits that stand alone are classified as outpatient services.

Event level matching is actually a series of matches between different categories of Medicare claims and MCBS service types. In conducting these matches, MCBS staff employ different match algorithms depending on the data elements available for the particular event categories being matched. The sequence of the matches is arranged so that the most similar MCBS event and Medicare claims categories are compared first (see Figure 2).

Each match algorithm employs a hierarchy of match criteria that are progressively less restrictive. For example, reported doctor visits are initially compared with claims data by doctor name, date of service, and total charge. If there is no exact match, the algorithm checks for a match on physician name and date of service or on total charge and date of service. If there is still no match, the program looks for an exact match on physician name and total charge with the date-of-service match relaxed to within a week. Thus, the match algorithms not only link a survey event and Medicare claim, but also indicate the strength of the link.

MCBS staff designed the match algorithms to allow survey-reported events to be linked to multiple Medicare claims and vice versa. There are several reasons for this. First, multiple links are often valid. For example, a survey-reported doctor visit may be linked to both a Medicare claim for physician services and a Medicare claim for lab services connected with the visit. Second, sometimes a stronger match occurs later in the series of matches than the initial, weak match. For example, a survey-reported doctor visit may have a weak match to a Medicare Part B physician/supplier claim and a strong link to a Medicare Part B outpatient claim. MCBS staff use the match strength indicator to resolve situations in which the multiple matches are logically inconsistent.

Figure 2. Overview of event category matches conducted during event level matching

Matches between similar service types
- IP to inpatient hospital
- MP, OM, SD, SL to Part B physician/supplier
- OP to outpatient hospital
- IU to SNF claims
- DU to Part B physician/supplier claims
- ER to outpatient hospital
- HF & HP to home health agency claims

Match between less similar service types
- ER to inpatient hospital claims
- OP to inpatient hospital claims
- IU to inpatient hospital claims
- IP to SNF claims
- IP to outpatient hospital claims
- OP to Part B physician/supplier claims
- MP, OM, SD, SL to outpatient hospital claims

Our strategy can be contrasted with a more probabilistic approach, such as that used by National Medical Expenditure Survey (NMES) for matching Medical Provider Survey data with household-reported data (Cohen & Carlson, 1994; Felligi & Sunter, 1969; Newcombe, 1988). Although many elements of the match process are comparable between the two surveys, for MCBS we did not assign a weight to the outcomes of the matching rules. Rather, the rules were arrayed in hierarchical fashion, reflecting the strength of the matches for each event category and across categories. Stronger matches were accepted before weaker matches for the same event.

A major concern in matching data from the two sources is potential double counting of medical events. MCBS staff have sought to minimize situations in which it is unclear whether an unmatched survey-reported event and an unmatched Medicare claim represent the same event or two different events. Such ambiguities were minimized by conducting the event level match within the data for each person. After organizing the data on a person basis, there are four possible outcomes: (a) a 100% match of the survey-reported events and Medicare claims; this does not present any reconciliation problems; (b) a 100% match of survey-reported events with unmatched Medicare claims; this does not present any reconciliation problems if we assume that the unmatched Medicare claims represent forgotten utilization additive to the sampled person's reported utilization; (c) a 100% match of Medicare claims with unmatched survey-reported events; this does not present any reconciliation problems if we assume that the unmatched survey-reported events are for non-Medicare services, unless the sampled person has reported that Medicare was a source of payment for the service; and (d) there are both unmatched Medicare claims and unmatched survey events; here there is a reconciliation problem.

MCBS staff attempt to address the fourth outcome by classifying unmatched survey events and unmatched claims into discrete service categories and determining whether the unmatched events and claims are in mutually exclusive categories. For example, an unmatched survey-reported dental visit and an unmatched Medicare inpatient hospital claim would be considered mutually exclusive and therefore classified as two separate events. The HCPCS[3] codes on the Medicare Part B physician/supplier claims are used to

[3]Codes that contain procedure specific information at several levels using the American Medical Association's Common Procedure Terminology (CPT) for physician services, HCFA codes for supplier services such as ambulance, and local codes that vary by carrier.

classify Medicare claims into a number of discreet subcategories. With this finer classification scheme, MCBS staff can be more precise in determining whether survey events and Medicare claims are mutually exclusive.

Event Level Match Results for 1992 Data

The first calendar year of MCBS utilization and expenditure data is 1992. Interviewers completed the collection of these data in August 1993. In June 1995, matching activities for most event types are essentially complete, and imputation activities for missing data are in progress. The post-matching file contains more than 300,000 events. Raw match results for the 1992 data by survey event type are presented for four major event classes in Table 1. Nearly one-half of the events are unmatched, and the proportion of false negatives is unknown. The difference between the minimum and maximum number of events is about 26% across these four event types, though it is only 11% for inpatient stays (which are among the most salient types of events for survey respondents) and it is 0% for hospital emergency room visits, since the Medicare system does not have that category as an event type in its own right.

Table 2 presents the results of our review of the unmatched claims and survey events at the person level to identify unmatched events, which must be nonduplicative (i.e., additive) because the individual did not have both unmatched survey events and unmatched claims. We were able to reduce the difference between the minimum and maximum number of events from 26.3% to 16.7% across these four event types.

It is informative to review the effect of the matching process on the expenditure data. For three event types, Table 3 presents the expenditure information as it looks after the match (but before imputation for missing data and editing for inconsistent data) by data source: administrative (i.e., Medicare claims) data or survey data. An event is classified as reported in both sources if it matches and has total charge (or Medicare-allowed charge) and at least some payment data from both sources. In the second group, an event is found in the administrative data that either does not match any survey event or that matches a survey event that has no reported dollars. In the third group, we see the opposite: a survey-reported event with dollars but either no matched event in the administrative data or a matched event with no dollars. The fourth group represents events for which dollars are missing from both sources.

For about 60% of the inpatient stays, expenditure data exist only in the administrative data. Most Medicare beneficiaries are unable to report any dollars associated with hospital stays that are covered by Medicare. For the other two event types shown in Table 3, medical provider visits and hospital outpatient department visits, about three-fourths

Table 1. MCBS raw match results

	A Matched survey-reported events	B Unmatched survey-reported events	C Unmatched claims	Maximum A + B + C	Minimum A + (B or C, whichever is greater)	Difference	
Hospital inpatient	2,853	1,474	493	4,820	4,327	493	(11.4%)
Medical provider	87,862	35,416	44,628	167,906	132,490	35,416	(26.7%)
Hospital outpatient	16,507	7,456	9,499	33,462	26,006	7,456	(28.7%)
Emergency room	1,160	1,030	—	2,190	2,190	0	(0.0%)
Total	108,382	45,376	54,620	208,378	165,013	43,365	(26.3%)

Table 2. MCBS match results after determining which nonmatches cannot be duplicates

	A Matched survey-reported events	B Non-duplicate survey-reported events	C Non-duplicate claims	D Unknown survey-reported events	E Unknown claims	Maximum A + B + C + D + E	Minimum A + B + C + (D or E, whichever is greater)	Difference	
Hospital inpatient	2,853	278	41	1,196	452	4,820	4,368	452	(10.3%)
Medical provider	87,862	11,254	3,009	24,162	41,619	167,906	143,744	24,162	(16.8%)
Hospital outpatient	16,507	2,311	537	5,145	8,962	33,462	28,317	5,145	(18.2%)
Emergency room	1,160	360	—	670	—	2,190	2,190	0	(0.0%)
Total	108,382	14,203	3,587	31,173	51,033	208,378	178,619	29,759	(16.7%)

Table 3. Preliminary distribution of source-of-expenditure data for three event categories

Group	Administrative data	Survey data	No. events	%
Hospital inpatient stays				
1	Reported	Reported	467	9.7
2	Reported	Missing	2,879	59.7
3	Missing	Reported	234	4.9
4	Missing	Missing	1,240	25.7
Total			4,820	100.0
Medical provider events				
1	Reported	Reported	74,505	44.4
2	Reported	Missing	57,985	34.5
3	Missing	Reported	13,302	7.9
4	Missing	Missing	22,114	13.2
Total			167,906	100.0
Hospital outpatient events				
1	Reported	Reported	9,843	29.4
2	Reported	Missing	16,163	48.3
3	Missing	Reported	2,771	8.3
4	Missing	Missing	4,685	14.0
Total			33,462	100.0

of the expenditure data is in the first two groups; that is, most of the events have dollars reported in both sources or in the administrative data alone. This reflects the dominance of the claims data in the MCBS design, even for those covered services for which many survey respondents are able to report expenditure data. The survey design focus is on amounts that are not covered by Medicare and on noncovered events.

It should be noted that Table 3 is based on preliminary data. Through additional editing and imputation, we expect some events will move from the top three groups into the fourth group and some events may move into different categories. However, even at this interim stage, the table shows how relatively dependent the MCBS is on administrative data (the Medicare claims) as opposed to survey data, at least for these three services that are covered by Medicare. In contrast, a similar analysis of the final data from the 1987 National Medical Expenditure Survey (NMES; a household-based survey that collected records from a sample of the medical providers reported by the household respondents and then matched these data to survey-reported events) showed a much higher proportion of total expenditure data reported by household respondents (Cohen & Carlson, 1994). This difference is expected, given the basic design differences between MCBS and NMES.

NMES reported the effects of the matching on estimates of total medical expenditures. We are unable to compare MCBS directly with NMES on this score, because the MCBS was not designed to produce independent estimates from administrative and survey data. However, we can compare (unweighted) data for the dollars on the average claim with dollars on the average survey report for the three event types. Table 4 shows that for hospital events (both inpatient and outpatient) in the first group (expenditures reported in both sources), the average survey report is much higher than the average claim. This reflects the effect of the Medicare program rules governing allowed charges for covered services. On the other hand, dollars for hospital stays reported by the survey respondent but not matched to a claim (the third group) are lower than the average claim amounts in the other groups. These inpatient stays may include a number of events that are more properly classified as outpatient services, including many surgical procedures.

Conclusions

Although matching survey data with Medicare data can introduce a number of ambiguities, the process improves estimates by increasing the amount of utilization and enhancing the accuracy of expenditure information. It reduces the need for imputation of missing data; through matching, we are able to supply total charges and at least some payment amounts by source for 86.4% of events in several major categories. Further research on MCBS match rates could be extraordinarily useful for informing decisions about optimal reference period lengths and for designing improved instruments, editing processes, and imputation strategies. We encourage future investigations of match rates by interviewer and respondent characteristics, proxy versus self-report, type of insurance coverage, length of panel experience, use of respondent records, Medicare claims service category, and Medicare fiscal agent.

Table 4. MCBS matching: Comparing dollars on Medicare claims and survey reports (unweighted data)

	Medicare dollars	MCBS dollars	No. events	Average $ claim	Average $ survey
Hospital inpatient stays					
	Reported	Reported	467	$6,508	$8,110
	Reported	Missing	2,386	$6,435	—
	No claim	Reported	234	—	$3,332
	No claim	Missing	1,240	—	—
	Reported	No survey-reported event	493	$5,833	—
Medical provider events: Reimbursement					
	Reported	Reported	74,505	$85	$89
	Reported	Missing	13,357	$71	—
	No Claim	Reported	13,302	—	$75
	No Claim	Missing	22,114	—	—
	Reported	No survey-reported event	44,628	$89	—
Hospital outpatient events: Reimbursement					
	Reported	Reported	9,843	$202	$353
	Reported	Missing	6,664	$201	—
	No claim	Reported	2,771	—	$139
	No claim	Missing	4,685	—	—
	Reported	No survey-reported event	9,499	$181	—

References

Cohen, S., & Carlson, B. (1994). A comparison of household and medical provider reported expenditures in the 1987 NMES. Journal of Official Statistics, 10, 3–29.

Felligi, I., & Sunter, A. (1969). A theory of record linkage. Journal of the American Statistical Association, 64, 1183–1210.

Newcombe, H. (1988). Handbook of record linkage: Methods for health and statistical studies, administration, and business. New York: Oxford Medical Publications.

Analytic Edits of SOP Values for Non-PM Events

Analytic editing of charge and source of payment data at the event level also determined some delta values. The general goal of the analytic edits was to resolve as many events as possible (i.e., to fully allocate total charges to payers) and to set as many delta values as possible based on logic and knowledge of how each payer operates. The edits resolved some events without using a hotdeck procedure to impute payment sources or amounts.

The analytic edits relied on having both unambiguous SOP values and external information about interaction among the insurance or payment sources. Edits for three of the nine sources (Medicaid, MCOs, and VA) depended on information specific to those payers, but delta values for other payment sources were also affected. The analytic edits are discussed next as they apply to each source of payment.

Medicaid: Analytic edits were used extensively when Medicaid was a potential or actual source of payment for an event. One set of edits--designed to reflect the role of Medicaid as the payer of last resort--ensured that Medicaid could not be a payer if payments were reported or imputed for another third-party insurer (except Medicare), or if the provider was a managed care organization (MCO) or VA facility. Another set of edits was developed for dual Medicaid/Medicare eligible beneficiaries whose cost-sharing liability is covered by Medicaid.

The following Medicaid edits ensured that Medicaid and another payer (except for Medicare and out-of-pocket) were never both sources of payment for the same event:

1. If private insurance, the VA, an MCO, or other private or public insurance (not Medicaid or Medicare) was a source of payment for an event, it was assumed that Medicaid was not also a payer (even if the respondent had reported a Medicaid payment) and the Medicaid delta component was set to 0.[1]

2. If Medicaid was reported as a definite payer for an event, all other payers with a delta value of missing were "turned off" as potential payers (set to 0).[2]

3. If the Medicaid delta value was missing (i.e., Medicaid was a potential but not definite payer for an event), and it was uncertain whether out-of-pocket, other public insurance, MCO, VA, or uncollected liability were sources of payment (i.e., their corresponding delta values were missing), it was assumed that Medicaid was a more likely payer and the delta values for the other payers were set to 0.

4. If, after the delta value imputations (described below), both private insurance and Medicaid were imputed as payers for an event, it was assumed that Medicaid was not a payer and its delta component was reset to 0.

Out-of-pocket payments were allowed when Medicaid was a payer only if the respondent was able to report the out-of-pocket amount. Medicaid usually picks up copays and deductibles for dual eligibles and Qualified Medicare Beneficiaries and the respondent has no out-of-pocket costs for Medicare-covered services.

Private and Medicare MCOs: MCOs (especially Medicare-contracting MCOs) often operate differently than other third-party payers and tend to have unique payment patterns. For instance, risk and (to a lesser extent) cost Medicare MCOs are paid a set fee per enrolled Medicare beneficiary (called a capitated amount) designed to compensate the MCO for the expected costs of delivering Medicare's package of benefits. There are no Medicare claims or Medicare or insurance statements indicating the total charge for events covered by the capitated amount. Often the respondent only knows the copay amount, if there was one. Also, MCOs often provide "Medigap"-type coverage by paying for most of the deductibles and copays for Medicare-covered benefits. A beneficiary who belongs to an MCO does not need private Medigap insurance or Medicaid coverage for these amounts. Thus, payment patterns for MCO beneficiaries tend to be simpler than those for fee-for-service beneficiaries. The set of analytic edits for MCOs attempted to account for these simplified patterns and for the respondent's usual inability to report charges and payments for events. The MCO edits also attempted to avoid creating "illogical" payment patterns.

1. If an MCO beneficiary reported a whole dollar total charge that was $15 or less, if the reported out-of-pocket amount equaled the reported total charge, and if there was no insurance statement, the reported total charge most likely represented only the beneficiary's out-of-pocket cost, not the full cost of the event. Therefore, the total charge was set to missing and imputed later in the program. In addition, the delta component for MCO was set to 1 and all other payers (except for out-of-pocket) were set to 0.

2. An SOP value of 3 for dental and medical provider events for MCO beneficiaries had a different interpretation than for other payers. MCO members were asked if the dental or medical provider service had been delivered by one of the MCO's providers or by an MCO-referred provider. If the answer to either of these questions was "yes," the MCO SOP value was set to 3 and the corresponding delta value was set to 1 instead of missing.

3. If an event occurred while the sample beneficiary belonged to a Medicare MCO, if the MCO was reported as a definite payer, and if there was no matching

Medicare claim and no insurance statement, all other payers (including Medicare) except out-of-pocket were assumed not to have paid for the event.[3]

4. If the MCO was a definite payer for an event, but information on the amount of the MCO's contribution or the total charge was unknown, other potential payers (excluding Medicare) with missing delta values were set to 0.[4]

5. If the MCO was a definite payer for the event, but information on the amount of the MCO's contribution or the total charge was unknown, other payers (including Medicare) with missing payment amounts were set to 0 even though the respondent reported them to be payers.

6. In some cases, the amount paid by the MCO was less than the total reported charge for an event and there were no other reported payment sources. For these events, one other payer's missing delta component was set to 1 to receive the residual dollars, in the following order: out-of-pocket, uncollected liability, Medigap insurance, private employer-sponsored insurance, other insurance, VA. Out-of-pocket was listed first as the most likely payer to have paid the remaining amount for an MCO event.

7. If the delta value for MCO was missing and if VA was a payer for the event or if there was an insurance statement, the MCO delta component was set to 0. It was assumed that the sample beneficiary's MCO would not be liable for any costs for VA-provided services. It was also assumed that if the respondent had a statement that did not indicate that the MCO paid for the service, the MCO most likely was not a payer.

Veterans' Administration (VA) coverage: If VA was a payer, no uncollected liability amounts were allowed. As both the insurer and provider of services, the VA does not "charge" more than it will be reimbursed by other payers. In this respect, services provided by the VA are similar to those provided by MCOs.

General Edits: At the beginning of the analytic editing, and after each main section of edits, an attempt was made to resolve events through addition or subtraction. Events without a known total charge but with a complete payment vector (i.e., each payer was identified as either having paid or not paid for an event and each payer's amount was known) were completed by summing across all payment sources to derive the total charge. Events with a known total charge and complete except for one missing payment amount or payment source were completed by subtraction. The excess of charges over known payment amounts was attributed to the known payer, or the one missing delta was set to 1 and the excess allocated to that payer.

If a service was provided free of charge, all delta values and payment amounts were set to 0.[5]

If a source was a potential payer for an event, or if the respondent reported that the payer had contributed to an event but did not know the amount, it was assumed that the payer was not actually a source if the current sum of reported payments equaled the reported total charge.

Notes:

1. The interaction of Medicaid and the category "uncollected liability" was handled slightly differently. If Medicaid were only a potential payer for an event but the SP had reported there was some uncollected liability, Medicaid was assumed not to have paid for the event. However, if the SP reported that Medicaid had paid for an event, it was assumed there was, in fact, no uncollected liability even if the SP had reported one. In many states, Medicaid payment rates are less than Medicare's and the state bases its copayment amounts on its own approved provider rates so that there is no "uncollected liability."

2. Medicare was not included in this edit since its delta value was never missing.

3. In these cases, it was also assumed that any total charges reported by the SP were probably not accurate since, without an insurance statement, Medicare HMO beneficiaries rarely know the total charge for an event. The total charge for the event was set to missing and imputed later in the program.

4. If the amount of the HMO's contribution or the total charge was not reported, other potential payers could be turned off without creating inconsistent payments and charges for the event.

5. If the event was reported as free, but the SP had also reported that a source other than Medicare or Medicaid had paid something for the event, the total charge was set to missing and imputed.

Setting SOP Flags

Each sample beneficiary's health insurance time line, survey-reported events and Medicare claims were used to establish an indicator variable (SOP flag) for each of the source of payment (SOP) categories. Information in the SOP flags was, in turn, used to determine the corresponding delta variables, which were used in imputation to determine whether or not a possible source of payment actually paid something toward the cost of an event.

This appendix outlines the rules that applied to the process of setting the values of the SOP flags. SOP flags can have one of five possible values:

0 = Source definitely did not pay
1 = Source reported as a payer, amount known
2 = Source reported as a payer, amount unknown
3 = Source possibly a payer, beneficiary was covered at the time of the event by applicable insurance
4 = Source possibly paid, but dates of insurance coverage, or of the event itself, are not clear

Setting initial values

SOP Medicare Medicare Part A and Part B entitlement dates established the period of Medicare coverage.

1. If the sample beneficiary was entitled to Medicare Part A benefits, Medicare was a potential source of payment for: Inpatient hospital -- "IP" events, SNF -- "IU" events and Home Health -- "HP" and "HF" events. The initial value of the Medicare SOP flag was 3 (possible payer) for these event types.

2. If the sample beneficiary was entitled to Medicare Part B benefits, Medicare was a potential source of payment for: Outpatient hospital -- "OP" events and Part B Physician/Supplier services -- "DU", "ER", "HP", "HF", "MP", "SD", "SL" and "OM" events. The initial value of the Medicare SOP flag was 3 (possible payer) for these event types.

SOP Medicaid If either the respondent or CMS administrative data indicated that the sample beneficiary had Medicaid coverage, the Medicaid SOP flag was initially set to 3 for all events which occurred during the period of Medicaid coverage.

SOP Managed care The managed care flag was set based on information in the beneficiary's health insurance time line and HCFA's administrative records of managed care enrollments.

1. If CMS administrative records indicated that the beneficiary was enrolled in a Medicare managed care plan but the beneficiary did not report the enrollment, the Managed care SOP flag was initialized to a value of 4 for all events that occurred during the beneficiary's enrollment.

2) We set the HMO SOP flag to 4, for all events except DU, MP and PM, if the Health insurance section shows that the SP was in an HMO, whether or not it is a Medicare HMO.

3. For DU and MP events where HMOASSOC and HMOREFER are applicable, if either HMOASSOC or HMOREFER = 1, the MCO SOP flag was set to 3 (possible payer, coverage definite); otherwise, if the respondent answered don't know (-7, -8 or -9) to either HMOASSOC or HMOREFER, the MCO SOP flag was set to 4 (possible payer, coverage not definite); else we set the MCO SOP flag to 0 (managed care organization did not pay).

4. For DU events, the MCO SOP flag was initialized to 3 if the respondent indicated that the MCO covers dental services, otherwise the MCO SOP flag was initialized to 4.

SOP Veterans Administration Information about the VA as a payment source was provided in the interview, by the respondent.

1. For all event types except prescription medicines, if the respondent indicated that the service was provided by a VA hospital or clinic, the VA SOP flag was set to 3; if the respondent was not certain that the service was provided by the VA, the VA SOP flag was set to 4; else the VA SOP flag was set to 0.

2. For drug events, the VA SOP flag was set to 4 if the VA paid a known amount for some other drug in the same round.

SOP Private health insurance - employer based Information about private health insurance as a payment source was provided in the insurance section of the interview, by the respondent, and through insurance statements used during the interviews. Information about the source of the policy (used to differentiate between employer-sponsored, and individually purchased private health insurance) was also provided by the respondent in the insurance section of the interview.

1. The employer-sponsored PHI SOP flag was set to 3 for all types of services, except prescribed medicines, which occurred while the sample beneficiary was covered by employer-sponsored health insurance, based on the health insurance time line and the date of the event.

2. For prescribed medicines, employer-sponsored health insurance was considered a possible source of payment (initial value SOP=3) if the respondent said that the plan covered drugs. If the respondent said that the plan did not cover drugs, but reported a specific amount the plan paid for another "PM" event, the employer-sponsored PHI SOP flag for all "PM" events during the same round was set to 4.

3. If the event date was missing or ambiguous and the sample beneficiary's insurance coverage changed during the round, the employer-sponsored PHI SOP flag was set to 4 instead of 3 where applicable.

SOP private health insurance--individually purchased Information about private health insurance as a payment source was provided in the insurance section of the interview, by the respondent, and through insurance statements. Information about the source of the policy (used to differentiate between employer-sponsored, and individually purchased private health insurance) was also provided by the respondent in the insurance section of the interview.

1. The Individually Purchased PHI SOP flag was set to 3 for all event types, except prescription medicines, which occurred while the sample beneficiary covered by individually purchased private health insurance, based on the beneficiary's health insurance time line and the date of the event.

2. For prescription medicines, the Individually Purchased PHI SOP flag was set to 3 if the respondent reported that the individually purchased PHI plan covered drugs. If the respondent said the plan did not cover drugs, but reported a specific amount the plan paid for another prescription medicine, the Individually Purchased PHI SOP flag was set to 4 for all prescription medicines reported in the same round.

3. If the event date was missing or ambiguous, and the sample beneficiary's insurance coverage changed during the round, the Individually Purchased PHI SOP flag was set to 4 instead of 3 where applicable.

SOP out-of-pocket and SOP uncollected liability

The out-of-pocket and uncollected liability flags were not set based on health insurance time lines. In many cases, these two categories could not be ruled out as payers based on the health insurance time line, or even after the claims match.

SOP other public insurance

1. For all events except prescription medicines, the Other Public SOP flag was set to 3 if the respondent reported coverage by "other public insurance".

2. For prescription medicines, the Other Public SOP flag was set to 4 if the SP reported that "other public insurance" paid a known amount for another medicine in the same round.

Updating SOP flags using survey-collected cost data

The initial values of the SOP flags were updated when survey-collected cost data provided more definitive information. If the respondent reported the amount the payer paid, the appropriate SOP flag was set to 1. If the respondent did not know how much the payer paid, the SOP flag was set to 2.

Updating SOP flags based upon matching Medicare claims data

The initial values of the SOP flags were also updated when the utilization could be linked to Medicare claims records.

Matched utilization and Medicare "claims only" utilization The Medicare payment amount and the Medicare SOP flag were updated if the survey-reported utilization matched Medicare claims data, or if the Medicare claims data provided the only record of the utilization. If the Medicare claims record showed a positive, non-zero Medicare payment, the Medicare SOP flag was set to 1, to show that the payment amount was known and would not have to be imputed. If the claims record showed that the sample beneficiary's Medicare benefits were exhausted, the Medicare SOP flag was set to 1, and the Medicare payment amount was set to $0.00. If the claims record indicated that the service was not a Medicare covered service or that the beneficiary did not have Medicare coverage for the service, both the Medicare payment amount and the Medicare SOP flag were set to zero.

If the claims record showed that Medicaid was a ssecondary payer, a Medicaid payment amount was imputed and the Medicaid SOP flag was set to 1.

If the claims record showed that Medicare was a secondary payer, the appropriate private insurance SOP flag was set to 1 (identifying the insurer as the primary payer), and the Medicare claim was used to develop the amount paid by the private insurer.

Unmatched "survey only" utilization The Medicare SOP flag was set to zero for all unmatched survey events unless the Medicare SOP flag already had a value of 1 or 2. If the beneficiary was a member of a managed care plan at the time of the event, the Medicare SOP flag was wet to 0 for any unmatched "survey only" utilization. The MCO SOP flag was not updated.