SOURCES AND SOLUTIONS FOR MISSING DATA IN THE NMCUES

Brenda G. Cox, Research Triangle Institute
Gordon Scott Bonham, University of Louisville

The rapidly rising cost of medical services in the United States in recent years, together with a continuous effort to improve the quality, effectiveness, and availability of health care, has lead to a continuing need for comprehensive data for individuals and families on health, health care, charges for care, and payers for care. The National Medical Care Utilization and Expenditure Survey (NMCUES), sponsored by the Health Care Financing Administration and the National Center for Health Statistics, was the second of a series of Federal surveys planned to provide this data on a regular basis. The survey permits in-depth statistical descriptions of the use and cost of health care services for the nation and for various population groups. It also provides valuable data for the evaluation of current public programs (such as Medicare and Medicaid), for the assessment of inequities in access to health care, and for the comparison of alternative health policy proposals.

The final NMCUES data needed to be in a form that would permit diverse types of analyses and would insure agreement among different users. The form of the data had to be such that aggregate national estimates could be obtained as well as accurate detailed relationships. Some standard handling of missing data was required in producing public use data files, or individual researchers would have had to make their own implicit or explicit decisions on how to handle missing data. This would result in different estimates from the same data. Some users, unaware of the problem of missing data, might arrive at unwarranted conclusions. For these reasons, procedures to compensate for missing survey data were an important part of developing the NMCUES data base. This paper describes the types of nonresponse found in the NMCUES and the procedures used to compensate for missing data in the public use data files.

The NMCUES, conducted by the Research Triangle Institute in conjunction with the National Opinion Research Center and SysteMetrics, Incorporated, had three major components. The National household survey component was based upon a stratified cluster sample of about 6,000 households representing the civilian, noninstitutionalized residents of the United States in 1980. Repeat interviews were conducted with the panel at approximately twelve-week intervals. In five rounds of data collection, information was collected on health, health care, health care cost, sources of payment, and insurance during calendar year 1980.

Four State Medicaid household surveys of Medicaid beneficiaries comprised the second component of the NMCUES. Administrative record data provided by California, Michigan, New York, and Texas were used to select a cluster list sample of Medicaid cases from each state. The selection procedures yielded aid-category balanced samples about of about 1,000 cooperating Medicaid cases per state [1]. The same instruments and data collection procedures were used as for the National household survey.

Administrative record data were extracted from Medicaid and Medicare files as the third component of the NMCUES. Medicaid eligibility during 1980 was collected for all people reported as covered in the household surveys. In addition, Medicaid claims data were extracted for people in the California, Michigan, New York and Texas State Medicaid household surveys. For older persons reported to be covered by Medicare in all the surveys, charge and payment data were obtained from the Federal Medicare files.

A number of field procedures were used to reduce the amount of missing data [2] Any adult member of the household could respond for other family members, reducing the necessity of contacting family members not at home at the time of the interview. This also meant that information about care prior to institutionalization or death was available for people who lived with others at the time. (Data were not collected for people who were institutionalized the whole year, or for times when a respondent was in an institution.) The interviews were conducted approximately three months apart to reduce memory loss. Boundary points of January 1, December 31, and previous interviews were used to reduce telescoping of events. Calendars were given to respondents to aid recall and to record health care events. The calendar also had a pocket to hold receipts or bills. A summary of previously reported events and cost was reviewed with the respondents during each interview. This review encouraged reporting of previously unknown information. Finally, an incentive was paid at the first, second and last interview to encourage cooperation and continued participation. Even with all of these procedures, however, there were still missing data.

1. SOURCES OF MISSING DATA IN THE NMCUES

Missing data result from a variety of sources, each having implications for the way the missing data should be treated. Data were not collected, and hence missing, when households could not be contacted. These noncontacted households include those whose members refused to be interviewed, those with no one available during the data collection period, and those whose household members were too sick to be interviewed or could not communicate. The people in these households are considered as "total nonrespondents." No information about them is available except that related to sample selection.

The second type of missing data is "partial nonresponse." Some people were initially interviewed, but data were not collected for the entire year. These people may have moved and not been subsequently located, have refused to participate after initially responding, or died or been institutionalized during the year with no one remaining to provide information for the period

prior to their death or institutionalization.

The third type of missing data is "item nonresponse." Individual items of data may be missing for a number of reasons: the respondent did not know, forgot, or refused to give the information; the interviewer failed to record the information; or the information did not get keyed into the data base. Item nonresponses for the NMCUES were generally characteristics of the person or characteristics of a health care event.

Missing data must be dealt with in all analyses. The most common way is to exclude cases with missing data from the analysis. Exclusion is a poor procedure since it ignores other information that is available for the data record. Table 1 illustrates. Considering each stage of possible missing data, 88.6 percent of the people in the original National sample frame provided data for the full time they were eligible during 1980 to provide data. People had an average of 4.5 medical visits during the year, and 77.6 percent of their medical visits had a known total charge. If people either knew charges for all or none of their visits, the analysis of the yearly medical visit cost to individual people could be based upon 68.8 percent of the original sample frame. If, however, visits with missing charges were randomly distributed among the population, analysis could be based upon 29.0 percent of the people in the original sample frame. Analysis of yearly cost for all care would be based on even a lower percentage.

Exclusion of missing data from analysis also requires the assumption that records with missing data are similar to those with known data. This is not the case for medical provider visit charges. Table 2 shows a greater percent of visits with unknown charges than visits with known charges had characteristics suggesting small or large charges for the visits rather than intermediate charges. There were also other differences to suggest that visits with unknown charges were not just like visits with known charges. It is best to use available information to assign a value where data are missing, or else to adjust the weights of survey respondents.

2. TREATMENT OF TOTAL NONRESPONDENTS

Total nonresponse is best handled through a weighting procedure. Only sampling related information is usually available for individuals who were never contacted. In the National household sample, known information was limited to characteristics of the geographical area in which the household lived. In the State Medicaid household samples, age, sex, race, type of Medicaid benefits received, and the number of people in the Medicaid case were also known.

The initial weight for each interviewed household was defined as the inverse of the household's overall selection probability. This initial weight was then adjusted to account for nonresponse and undercoverage [3,4]. Although 80-97 percent of the eligible households were interviewed (varying by sample), a biasing effect on survey estimates of means and proportions could result if nonresponding households had different health and health care experiences than responding households. Further, national totals would be underestimated unless some allowance was made for the loss of data due to nonresponse. Weights for responding households were increased to account for nonresponse in the same sampling unit. Post stratification of the survey estimates were then made to official figures. Post stratification adjusted for nonresponse and undercoverage that could have occurred at different rates for different groups of people, defined by age, race, and sex.

Some information was obtained in the Round 1 interview for individuals who were still eligible but not subsequently interviewed (0.3 - 1.0 percent of all individuals). The question arose: should they be treated as total nonrespondents or as respondents with only partial information? It was decided that these individuals had so little health care data that they should be treated as total nonrespondents. Operationally, total nonrespondents were defined as those individuals with data for less than 1/3 of the time they were presumed eligible for interview (i.e., part of the civilian, noninstitutional population of the United States). Data that had been obtained for a total nonrespondent were ignored.

A situation exists in panel or longitudinal data that should not be confused with missing data. Some people are in the sample universe for only part of the time. These individuals present problems for analyses that classify individuals by annual totals, such as the amount of medical care expenses during the year. An individual in NMCUES who incurred very high monthly medical expenses would appear to have had low yearly medical expenses if he or she only lived one month. This individual had no data for eleven months of the year, but the data were not missing. Data were not applicable for months in which individuals were ineligible for the survey. For data analyses, an adjustment can be made that uses the proportion of the year that the person was eligible to provide data. Then, for certain analyses, a yearly rate of medical expense can be calculated which accounts for the time that individuals were members of the civilian, noninstitutional population. No imputations were needed for time periods when people were ineligible for interview.

3. TREATMENT OF PARTIAL NONRESPONSE

Over the five rounds of data collection, approximately two percent of the people in the initial National household sample were lost through attrition. There were an additional four percent that were not lost, but that had gaps in their data during times they were eligible. The equivalent figures for the State Medicaid household sample were four to six percent lost and an additional three to eight percent with data gaps. Table 3 for the National household sample shows the variation in the level of missing data with characteristics of respondents. If people with higher rates of missing data were also those with

more medical visits, the number of medical care events would be greatly underrepresented without attrition imputation.

The variables shown in Table 3 were though relevant to the number of medical events that would be missed for a person with partial nonresponse, and hence candidates for classing variables. In forming imputation classes, the overall goal is to form classes for which responses are homogeneous within each class, heterogeneous between classes, and for which the rate of missing data varies. Further, the characteristics used to define the classes have to be known for both respondents and nonrespondents.

Two types of data were missing for partial respondents. One type related to characteristics that did not change during the year, but were measured during a round when the person was not interviewed. These missing data were handled as item nonresponse. The other type of missing data related to health care events that occurred during the time that data were missing, and were related to the length of time the person was not a respondent. Partial nonresponse for health care events was accounted for through attrition imputation. The match with administrative records was used as a complement or alternative to attrition imputation in the State Medicaid surveys and is discussed separately.

Attrition imputations were made using a weighted hot deck imputation procedure [5]. The imputation occurs within imputation classes so that the distribution of means and proportions is preserved within each class over repeated imputations. This imputation strategy may be thought of as utilizing two data files, a data file of respondents (donors) and a data file of nonrespondents (recipients). Data for responding individuals are substituted for missing data for nonresponding individuals. The first step is to sort the two data files with respect to person characteristics (classes) related to response and the items of interest. Both files have sample weights attached to each individual. The number of times that the data for a donor is accessed to impute to recipients is defined as a function of the sampling weight for the donor and of the recipients to which the information can potentially be imputed. For time periods for which the recipient had data missing, the visits (if any) reported by the donor for the same time period were imputed to the recipient. If the donor was ineligible to provide data for any part of the time period for which the recipient had data missing, the recipient was imputed to be ineligible during the same time period.

In the NMCUES, attrition imputation accounted for 3.1 to 6.9 percent of the event records (Table 4). Since attrition imputation attributed the total donor event record to the recipient, certain uses of the imputed records are appropriate and other uses are not appropriate. Imputed records should be used in classifying a person by the number of medical visits made during the year. They should also be used in estimating the amount of care and the cost of care in 1980 associated with certain conditions, e.g., having a baby. However, it is not appropriate to categorize a person as pregnant when a prenatal care visit record was imputed. The conditions associated with the individual were not used as classing variables in the imputation process.

4. TREATMENT OF ITEM NONRESPONSE

Data could be missing for a respondent even though the respondent was interviewed in all data collection rounds. Additionally, data could be missing because the respondent was not interviewed in a particular round in which the data were obtained. Individual items about a doctor visit, hospital stay, or prescribed medicine could also be missing. Item imputations were made in three ways: logical, simple hot deck, and weighted hot deck [6]. Use of the administrative record match could be considered as a fourth way of item imputation, but is discussed separately.

Logical imputations were used whenever similar information were available in the record. As an example, when the sex of the person was missing, the person's relationship to the head of the household was checked to determine if relationship was gender-specific, e.g. "wife" or "son." If so, the sex variable was logically imputed based upon the relation to head.

Simple hot deck imputation was used when the amount of missing data was small, the item was not a major analytic variable, or the item (age, race, and sex) had to be imputed prior to weight construction. For the simple hot deck imputation procedure, respondents are divided into imputation classes by characteristics related to the item being imputed. Within each class, the records are generally sorted by variables related to the item being imputed. An initial value is determined for each class based upon previous or current data. As the new data are processed, the imputation class to which each individual belongs is determined. If the record being processed is complete, then that record's response is supplied for the cell of the hot deck. When a record is encountered with a missing item, the response in the cell of the hot deck is imputed for the missing response.

Weighted hot deck imputation was used in the NMCUES whenever a large number of records had missing data or the missing data were key analytic items. The procedure was the same as described for the attrition imputation except single items were imputed rather than entire records. The procedure is designed so that within imputation classes the means and proportions estimated from the imputation-revised data will be equal in expectation to the means and proportions estimated using only complete respondent data. Variances, covariances, correlations, regression coefficients, and other higher order population parameters estimated from the imputation-revised data will also equal the corresponding estimator obtained from the respondent data alone.

It was not feasible to replace missing data for all data items because of the size and

complexity of the NMCUES data base. There were about 1,400 data items for each of the 36,000 people included in the surveys. The NMCUES approach was to designate about five percent of the data items as important enough to merit missing data imputations. Items for which imputations were made cover the areas shown in Table 5. These items were the most important variables for analysis.

The items subject to imputation were divided into sets and imputations performed within those sets. For each set of data items, the most cost effective imputation strategy consistent with quality requirements was selected. By reducing the number of passes through the large NMCUES files, some approaches could drastically reduced data processing costs while producing results that were essentially comparable in quality to other approaches.

Imputations were conducted independently within the five National and State samples.

5. ADMINISTRATIVE RECORDS

Data were collected from the State Medicaid household samples in the same way as for the National household sample. However, a much higher rate of missing data was expected, and encountered, for health care charges than in the National sample. Medicaid pays the total bill directly to the provider, and the beneficiary seldom has any knowledge of the charge. Because four-fifths or more of the data were missing, total charges in the State Medicaid samples were not inputed from other survey-reported visits. Rather information was used from Medicaid claims records. Matching of claims for hospital stays, doctor visits, and other medical expenses with household records was done by coders using a written set of procedures. Once matched, charges and payments recorded in the claims data could be used in place of missing household data.

For the Texas Medicaid household sample, there was about an 85 percent agreement among three independent judges on what constituted a match between household reported events and Medicaid record events. Respondent name, visit date, provider name and address, type of visit, and sources of payment were used as matching criteria. All events except for prescribed medicines were matched. Unmatched Medicaid claims records provided additional health care events much like the attrition imputation. Unmatched claims records were counted as actual visits. To keep from double counting, unmatched household reported or inputed visits of the type and during periods covered by claims records were not counted as visits for the best estimate files.

Charge and some source of payment data from the administrative claims record were available for most visits with missing household charge data. For the small percentage of medical events in the State Medicaid household samples that had reported charge data, the household data were used to amplify claims data. For those medical care events with both missing household and claims data, weighted hot deck imputation was used in a manner similar to that described for the National household sample.

The Medicare administrative records were also matched and linked to household-reported data in the National and State Medicaid surveys. Medicare administrative data was used as a correction or amplification of information more than as an alternative for imputation. First, many people did know their total charges for health care which was only partially covered by Medicare, and there was not as high a rate of missing data as for Medicaid-covered care. Second, only charges for hospital stays could be associated with individual health care events given the structure of the Medicare administrative records. Therefore, the Medicare records mostly provided aggregate information at the person level rather than at the event level.

6. CONCLUSION

Missing data is present in any survey, regardless of the care and quality with which it was conducted. Ignoring missing data can severely restrict the number of records available for analysis, and produce biased results. Ad hoc adjustments for different analyses of the same data can produce different figures for what should be the same estimate. Careful imputation permits use of related information with the minimal amount of assumptions and assures that statistics based upon imputed data have the same properties as those based upon collected data.

The National Medical Care Utilization and Expenditure Survey was conducted in 1980 to provide data on health, health care, and health care expenses. Imputations were made for missing data in these basic analytic areas, supplemented by a few imputations for characteristics of respondents. Imputations for total nonresponse were made through a weighting strategy. Imputations for partial nonresponse were made through a weighted hot deck imputation procedure. Imputations for item nonresponse were made using logical imputations, simple hot deck imputation and weighted hot deck imputation procedures.

Medicaid and Medicare administrative data were also obtained as part of the NMCUES. Medicaid claims data were used as the primary means to compensate for missing health care charges in the State Medicaid household surveys. The level of missing data for care covered by Medicaid was so high that any statistical imputation strategy would be highly variable and probably biased. However, even administrative record data is not complete, and some imputation was still necessary in producing best estimates.

There are many different ways to handle missing data. All take time, have cost, and have analytic implications. The worst procedure, however, is to ignore missing data. The development of the NMCUES data base incorporated a wide range of strategies to handle missing data. The result is a valuable data base useful for many different types of analysis.

REFERENCES

[1] Folsom, R. and Iannacchione, V., "NMCUES State Medicaid household Survey Sample Design Statement." Contract No. HRA-233-79-2032. Research Triangle Park, NC: Research Triangle Institute, February 1980.

[2] National Center for Health Statistics, Bonham, G., "Procedures and questionnaires of the National Medical Care Utilization and Expenditure Survey," Series A, Methodological Report No. 1, DHHS Pub. No. 83-20001. Washington: U.S. Government Printing Office, March 1983.

[3] Jones, B., "Development of Sample Weights for the National Household Component of the National Medical Core Utilization and Expenditure Survey." Contract No. HRA-233-79-2032. Research Triangle Park, NC: Research Triangle Institute, April 1982.

[4] Iannacchione, V., Cox, B. and Folsom, R., "NMCUES State Medicaid Household Surveys Weighting Methodology." Contract No. HRA-233-79-2032. Research Triangle Park, NC: Research Triangle Institute, April 1982.

[5] Cox, B., "The Weighted Sequential Hot Deck Imputation Procedure," Proceedings of the American Statistical Association, Survey Research Methods Section, 721-726, 1980.

[6] Cox, B., Parker, A., Sweetland, S. and Wheeless, S., "Imputation of Missing Item Data for the National Medical Care Utilization and Expenditure Survey." Contract No. HRA-233-79-2032. Research Triangle Park, NC: Research Triangle Institute, June 1982.

TABLE 1

Proportion of National household sample from population estimated to have completely reported charges for medical visits

Round 1 reporting unit response rate		.911
Round 2 person response rate		.996
Complete year person response rate		.977
Complete respondents		.886
	Clustered	Random
Total charges response rate if missing charges are clustered among people or randomly distributed		
(average 4.5 visits with .776 having known charge)	.776	.327
Proportion of sample persons with complete yearly charges	.688	.290

TABLE 2

Percent of medical visit records by post imputation total charge, according to whether the charge was known or unknown: NMCUES national household sample

Post imputation charge	Known charge	Unknown charge	Difference
All	100.0	100.0	—
No charge	10.5	12.4	+1.9
$0.01 - $10.00	15.8	17.4	+1.6
$10.01 - $17.00	22.9	22.1	-0.8
$17.01 - $30.00	27.0	24.8	-2.2
$30.01 - $86.00	19.0	18.2	-0.8
$86.01 or more	4.8	5.1	+0.3
(Number of records)	(59,390)	(17,073)	—

Table 3: Annual Complete Data Rates for Key Individuals From Responding Round 1 Reporting Units and Average Number of Medical Visits, by Selected Characteristics

Selected Characteristics		Annual Complete Data Rate	Average Number of Medical Visits
Total		94.0	5.11
Age of Individual	0-16	94.7	3.73
	17-29	93.1	4.43
	30-44	94.7	4.92
	45-54	94.4	5.54
	55-64	94.5	6.48
	65+	92.7	7.96
Race of Individual	Black	92.2	3.90
	NonBlack	94.3	5.26
Sex of Individual	Male	93.9	4.32
	Female	94.1	5.83
Education of Head of Household	0	82.7	6.57
	1-8	94.4	5.22
	9-12	93.9	4.83
	13+	94.7	5.42
Number of Medical Visits in First Quarter	0	93.2	1.98
	1	95.1	4.18
	2	94.4	6.53
	3-4	95.6	9.53
	5-6	95.9	14.56
	7-8	94.4	19.02
	9+	93.6	38.45
Self-Reported Health Status	Excellent	95.1	3.64
	Good	93.9	5.27
	Fair	92.9	9.05
	Poor	91.0	13.54
Health Plan	Medicare	92.9	8.75
	Other Public	92.9	5.45
	Private	95.3	4.59
	Uninsured	87.2	2.54

Table 4: Results of Attrition Imputation By Type of Events

Status	Dental Visits	Hospital Stay	Medical Visit	Pres. and Other Exp.
Total Records	35,703	7,456	185,386	121,180
Original	33,251	7,026	179,713	116,928
Imputed	2,452	430	5,673	4,252
Percent Imputed	6.9	5.8	3.1	3.5

TABLE 5

Items imputed in the NMCUES by percent missing data and type of imputation

Item	Percent Missing	Type of imputation			
		Logic	Hot-deck	Weighted	Records
Age and birthdate	0.2 - 1.1	x	x		
Race/Hispanic origin	22.3 - 22.6[a]	x	x		
Sex	0.5	x	x		
Education level	1.6	x	x		
Employment status	1.5 - 3.3			x	
Disability days	28.7 - 48.9			x	
Nights hospitalized	5.3 - 9.0	x	x		
Health insurance premium	12.5 - 24.5			x	
Income	4.2 - 43.3	x		x	
Health care charges and payments	14.0 - 90.0[b]			x	x

[a] Race and Hispanic origin were logically imputed for all children based on the race and origin of adult household members.

[b] Estimated magnitude of missing data in State Medicaid household surveys.