IMPUTATION IN SAMPLE SURVEYS: THE EFFECT ON SMALL DOMAIN ESTIMATES

Mary Grace Kovar, National Center for Health Statistics

It is an unfortunate characteristic of surveys (and of censuses) that the respondents sometimes will not or cannot tell us the answers to the questions we ask. At times it is useful to know that the respondent does not know the information and we leave the response as an unknown. At other times it is essential that we have the information for every item.

The National Medical Care and Expenditure Survey, the NMCUES, was a case where it was essential to have information for each item. A major purpose of the survey was to estimate the total medical care utilization of the United States civilian noninstitutionalized population in 1980 and the total expenditures for the care. Failure to assign values for missing data would result in underestimates. If, for example, the respondent knew that someone in the family had visited a doctor but did not know how many times or what the visits cost, the estimated TOTAL charges would be biased downward unless values were imputed for the number of visits and the charges for each.

A great deal of detailed financial information was required. Each contact for medical care and the charges for that care by characteristics of the person, the medical-care provider, and the sources of payment was essential. Individual income by source and family income were needed to study income, health insurance coverage, utilization, and expenditures. Not surprisingly, a rather large percentage of the income and charge items had to be imputed. The percentages are in the documentation of the public-use data tapes (NCHS, 1984). Values for selected items are shown in Table 1.

However, estimating such individual items was only the first step. Summary measures were needed. To obtain the total income for the person, it was necessary to sum over all sources of income. To obtain the total charge for, say, a hospitalization, it was necessary to sum over the individual components. When the charge for any one of the components was imputed, the total charge for that episode was considered to be imputed. Thus, the percentages of the total income of a person or the charge for an episode of medical care that were imputed are higher than the percentages for the individual items. Selected values, as given in the data-tape documentation, are shown in Table 2.

The level of imputation for the total family income, the total charges for ALL medical care, or the total from any given source of payment, such as private health insurance, is obviously even higher. These totals may be the sum of dozens of individual components, and many or all of the components for a person or a family may have been imputed.

The imputation can create analytic problems for several reasons. As you have seen, the proportion of some of the items for which imputation was needed was large. Furthermore, one of the characteristics of respondents is that if they do not know one piece of information, they are likely not to know many others. Such a respondent, or even worse the family of that respondent, will have multiple imputed pieces of information. Relationships among variables may be obscured. And finally, many of the domains of analytic interest were not the characteristics used for imputation.

The NMCUES was relatively small for a national survey; there were approximately 6,000 reporting units containing 17,123 people in the national sample. Defining "similar" so that there was an adequate number of respondents in each imputation cell presented problems. The number of characteristics used for imputation had to be limited. Even then it was necessary to collapse cells for some population groups and some records were used as "donors" more than once.

Hospital care is by far the most expensive form of medical care and consumes the largest proportion of the medical care dollar. Most of the proposals for containing the cost of medical care have focused on hospital care - the use of diagnostic related groups for hospital reimbursement and the substitution of hospice care for inpatient hospital care are current examples. Because of the importance of the charges for hospital care, the remainder of this paper will be devoted to the effect of imputation on those charges for two types of small domains - age and diagnostic groups.

The imputation was done by weighted hot deck procedures (Cox and Folsom, 1981). The imputation had to be done in stages to assure that the data needed for imputing the charge data were available. In essence, after everything else had been imputed the following scheme was used (Cox et al, 1982):

Classification 1	Classification 2	Sort
Born in 1980	work done	work done
Not born in 1980		
Delivery	nights	age
Operation	nights and work done	age
Other	nights and work done	age

Work done is a variable created by counting how many X-rays, laboratory tests, and diagnostic procedures were done.

The analytic impact of the imputation is evaluated first by comparing means based on real (reported) data, imputed data, and the survey data. Data for this comparison were imputed by the weighted sequential hot deck procedure, the procedure actually used to impute hospital charges in the NMCUES (Cox, Sweetland, and Wheeless 1982). Two estimates of the total charges for hospital care are then compared --one computed from $\hat{N}\bar{y}_s$ (where $\bar{y}_s$ is the survey mean) and the other computed from $\hat{N}\bar{y}_r$ (where $\bar{y}_r$ is the mean based on the real data. All analyses were done from the public-use data tape and can be replicated or expanded by anyone who has the tapes. The imputed values are clearly flagged on the tapes. On the hospital and medical visit files the imputation indicator shows whether the item was real or imputed and, if real, whether it was used as a donor -that is, whether it was used to impute data to other records.

Excluding the birth episode of newborns, there were 2,710 in-scope hospital discharges. Of these, 1,709 (63 percent) had real data on the total charge for that episode and 1,001 (37 percent) had the total charge imputed (Table 3). The mean for the former is $2,191 and for the latter $2,322. The overall mean used for analysis, the survey mean based on real and imputed data, is $2,240. The difference is small - only $49. However, the 2,710 discharges yield a national estimate of 35,700 thousand discharges in 1980. If the total charge for hospitalization that year had been estimated by assuming that charges for those for whom data were missing were the same as those for whom we had information ($\bar{y}_m = \bar{y}_r$), the estimated total for the country would have been $1.7 billion less than the estimate we are using.

The effect of the weighted sequential hot deck imputation, then, was to increase the estimate of the total amount spent for hospitalization in 1980 over the estimate based on the assumption that the mean charge for missing data was the same as that for real data. One reason this happened was that the charges for "donors" were higher than the charges for the "non-donors".

However, that is not the case for all of the individual age or diagnostic groups. The survey mean is lower than the real-data mean for 7 to the 11 age groups and 11 of the 19 diagnostic groups (Tables 3 and 4). In some cases the differences are substantial. For example, the survey mean is $540 higher for 18-24 year olds and $216 lower for 25-34 year olds than the real-data mean. The survey mean is $505 higher for a urinary condition and $664 lower for a malignant neoplasm than the real-data mean. In contrast, and I will return to this later, the difference between the survey and the real-data mean is only $65 when the hospitalization was for a circulatory condition.

These examples were not chosen from among extremely small age groups or rare conditions selected because hospitalizations were unusual events and subject to a great deal of variation. The differences would have a major impact on conclusions. The survey estimate of the national hospital bill for malignant neoplasms is $7.6 million. It would be 13 percent higher -$8.6 billion - a difference of a billion dollars, if estimated from the mean for real data (Table 5). Similarly, the national estimate for respiratory conditions would be $8.0 billion instead of $6.7 billion - a difference of $1.3 billion - if estimated from the mean of real data. In the other direction, the bill for urinary conditions would be $5.3 billion instead of $7.0 billion - a difference of $1.7 billion.

In addition, the ranking of diagnosis by charge per hospitalization and the attribution of the proportion of the total cost of hospitalization accounted for by each of the diagnostic categories are affected by the choice of imputation methods.

There is no consistent pattern. The mean for diagnoses with high real-data average charges is not consistently lowered by imputation nor is the mean for diagnoses with low average charges consistently raised. Diagnostic categories with a small number of cases are not necessarily affected more than those with a larger number.

Other relationships of analytic interest are also affected. There is, for example, a great deal of interest in how the medical care of poor people is paid for. The proportion of hospital charges imputed varied inversely with family income from about 71 percent of the hospitalizations of people in families with 1980 incomes of under $3,000 to 25 percent for people in families with incomes of $25,000 or more.

The survey estimate of the mean charge per hospitalization if a person in the lowest income category is $3,584 - $596 higher than the reported mean charge. The difference for the highest income category is only $4.

Why are the imputed charges so different from the "real" charges and what is the analyst to do?

The basic problem is that people do not know the total charges for their medical care. There are a lot of reasons. One common reason is that public or private health insurance pays a large portion of the bill and the patient may not know what the total charge was. It depends upon how the billing was done and what was sent to the patient. A second reason is that the person may have been hospitalized in a public hospital and the total charge is the charge for incidentals. The patient honestly reports the total that he or she is aware of. Public Health Service hospitals, Veterans' Administration hospitals, and many municipal hospitals do not charge the patient most of the cost and the patient does not know what the hospitalization really cost. The individual is reporting honestly but the true charge is unknown to him or her. Many enrollees of HMO's may not know the true charge because the enrollee simply pays a per capitation charge. Thus, even the real data may not measure what we intended to measure.

With the benefit of hindsight we can see that hospital ownership should have been used as an imputation category (Table 6). The mean charge reported for hospitalizations in Federal hospitals was $6. But 76 percent of the hospitalizations in Federal hospitals has unknown charges. Data were imputed from other records for those hospitalizations and, because the donors were records of hospitalizations in other kinds of hospitals, the imputed values were much higher, and the survey mean is $3420.

The choice of age as a sorting variable may have caused some of the differences. Relatively few of the charges for young adults had to be imputed. Relatively large proportions of their records had real data that could be used to impute data for older adults whose records would follow.

Newborns were a class regardless of whether the hospitalization was for the birth or was a separate hospitalization of the infant. Most hospitals do not have a separate birth-episode charge for the infant; the total charge is on the mother's bill. Those infants with zero charges were used to impute charges for separate admissions of infants. Thus, the imputed charges for congenital defects are much lower than the real charges.

Some errors were missed. There were six hospitalizations with charges of $90,000 or more -- three real and three imputed. The three that were real were all coding errors. The imputation of $100,000 to an 18-year-old with a urinary condition certainly contributed to the mean imputed charge for urinary conditions being twice the mean based on real data. The imputation of $117,155 to an 82-year-old woman with a fracture certainly contributed to the high imputed mean for external causes.

I certainly do not recommend that an analyist follow the procedure used here and use the means based on real data for several reasons. First, that would require the unwarrented assumption that charges for hospitalizations where the charge was not reported were the same as reported charges. There is ample evidence that they were not. Second, each analyst would be using different means depending on the category of interest. Results would differ from one analyst to another and readers of the publications would be confused. Finally, the variances of the estimates would be reduced. They would be greatly reduced in categories or for variables where there was a great deal of missing data because a large portion of the values would be set at the mean value.

The data should be used with the "survey" imputations that preserve the distributions. However, the analyst who is using the data does have a responsibility to examine the data including the imputed values that are flagged on the NMCUES public-use tapes.

The lack of consistency in the impact of imputation means that analysts interested in particular problems

must investigate the distributions that are relevant to their problems very carefully before proceeding. The information in this paper is not sufficient for all decisions. For example, as noted, the mean charge for hospitalization for circulatory conditions is relatively unaffected by the imputation method. That does not mean that all of the subcategories are equally unaffected. The survey mean is $284 higher than the real-data mean for ischemic heart disease and $901 lower for cerebrovascular disease ($1,262 vs $978 and $3,107 vs $4,008).

Outliers should be examined. Imputed values should be examined. The analyst can reimpute if that seems appropriate, and document what was done. (Kovar 1983). Examination of the type of hospital, sources of payment, and kind of surgery can provide additional information to use in interpretation.

And finally, the analyst should be careful not to overinterpret the data. There is fascinating and useful information from the NMCUES and the level of imputation is much lower for other data than for charges. It should be used but used with care. Read the publications on the procedures and questionnaire and the documentation of the public-use data tapes very carefully (Bonham 1983, NCHS 1984). That is why they were published.

References

1. Bonham, GS: Procedures and questionnaires of the National Medical Care Utilization and Expenditure Survey. National Medical Care Utilization and Expenditure Survey. Series A, Methodological Report No. 1. DHHS. Pub. No. 83-20001. National Center for Health Statistics, Public Health Service. Washington. U.S. Government Printing Office, Mar. 1983.
2. Cox, BG, AE Parker, SS Sweetland, and SC Wheeless: Imputation of missing item data for the National Medical Care Utilization and Expenditure Survey. Research Triangle Institute. Research Triangle Park, NC., June 1982.
3. Cox, BG and RE Folsom: An evaluation of weighted hot deck imputation for unreported health care visits. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1981.
4. Kovar, MG: Expenditures for the medical care of elderly people living in the community throughout 1980. National Medical Care Utilization and Expenditure Survey, Data Report No. 4. DHHS Pub. No. (PHS) 84-20000. National Center for Health Statistics, Public Health Service. Washington. U.S. Government Printing Office, Nov. 1983.
5. National Center for Health Statistics: Public use data tape documentation. National Medical Care Utilization and Expenditure Survey, 1980. Public Health Service. Hyattsville, Md., Apr. 1984.

Table 1. Percentage of selected data items revised through imputation. National Medical Care Utilization and Expenditure Survey, 1980.

Data Item	Percent Imputed
Sources of income (17,123 records)	
Income from wages, salary, etc.	9.7
Income from interest	21.6
Income from investment	6.4
Charges for medical visits (86,594 records)	
Amount from first source of payment	11.6
Amount from second source of payment	7.0
Amount from third source of payment	2.1
Charges for hospital care (2,946 records)	
Amount from first source of payment	17.6
Amount from second source of payment	16.2
Amount from third source of payment	9.5
Charges from first doctor in hospital (2,946 records)	
Amount from first source of payment	12.6
Amount from second source of payment	10.9
Amount from third source of payment	2.7
Charges for dental care (23,113 records)	
Amount from first source of payment	6.9
Amount from second source of payment	5.2
Amount from third source of payment	2.9
Charges for prescribed medicines and other expenses (58,544)	
Amount from first source of payment	10.0
Amount from second source of payment	6.8
Amount from third source of payment	1.4

NOTES: There were eleven possible sources of income. Although every medical record had to have a first source of payment, fewer than half had a second source and very few had three or more.

Table 2. Percentage of selected totals revised through imputation. National Medical Care Utilization and Expenditure Survey, 1980.

Selected Totals	Percent Imputed
Personal Income	30.4
Charges for medical visits	25.9
Charges for hospitalization	36.3
Charges for first doctor in hospital	15.8
Charges for dental care	13.8
Charges for prescribed medicine, etc.	19.4

Table 3. Hospital Charges by Age and Imputation. National Medical Care Utilization and Expenditure Survey, 1980.

			Imputation Status			
	Total	Imputed	Real	Not donor	Donor once	Donor twice
			Estimated Mean Charges			
Total	**$2239.78**	**$2321.61**	**$2191.12**	**$1926.78**	**$2373.49**	**$ 2561.05**
Newborn	1819.32	1565.79	2007.99	873.44	2691.33	0.00
Under 6	1281.08	1270.29	1292.23	853.13	1568.57	921.73
6-17 Years	1134.61	871.42	1316.07	1425.56	1260.78	391.66
18-24 Years	1920.99	2991.41	1381.24	1245.95	1541.25	964.22
25-34 Years	2018.39	1506.29	2234.11	1670.85	2790.24	2774.24
35-44 Years	1919.48	1985.66	1892.14	1661.41	2036.80	1983.69
45-54 Years	1915.34	1866.00	1944.51	2103.88	1841.55	1533.16
55-64 Years	3075.25	3477.33	2838.08	2420.86	3114.69	3147.10
65-74 Years	2361.87	2283.81	2411.30	2490.13	2408.80	312.70
75-84 Years	3777.48	3584.62	3942.82	3798.84	3725.96	12471.13
85 Years Plus	2599.21	2774.78	2350.18	2118.47	2260.99	5394.00

NOTES: The birth episodes for newborns are excluded.

Table 4. Hospital Charges by Diagnostic Group and Imputation Status. National Medical Care Utilization and Expenditure Survey, 1980.

Diagnostic Group	Total	Imputed	Imputation Status: Real			
			All	Not donor	Donor once	Donor twice
	Estimated Mean Charges					
Total	**$2239.78**	**$2321.61**	**$2191.12**	**$1926.78**	**$2373.49**	**$2561.05**
Infectious	1322.31	1847.34	1073.27	952.82	1131.67	1292.00
Malignant neoplasms	4922.28	3879.15	5586.40	3949.48	6237.22	13218.40
Other neoplasms	1763.04	1511.29	1865.67	1963.32	1679.30	1956.00
Metabolic	2072.86	2256.01	1958.80	1956.05	1976.45	1827.68
Blood	1719.37	1100.60	1946.65	1269.16	2423.30	
Mental disorders	2057.55	1624.15	2456.87	1989.87	2591.67	2697.61
Nervous/sense	1484.94	1338.88	1550.41	1301.06	1772.31	662.00
Circulatory	2692.09	2765.26	2627.09	3055.77	2301.45	7205.91
Respiratory	2007.01	1418.75	2378.23	2134.82	2589.64	891.15
Digestive	2352.37	2216.36	2417.08	1941.06	2897.34	868.00
Urinary	2093.51	3326.58	1588.64	1476.02	1699.00	1200.37
Pregnancy	1281.64	1247.95	1297.69	1421.26	1137.41	108.00
Skin	1279.96	1085.00	1386.37	784.52	1318.00	5394.00
Muscle	2506.14	2499.63	2509.67	2387.87	2496.53	091.89
Congenital	3202.20	289.20	4254.18	5122.20	3816.39	
Symptoms	2486.01	3066.86	1878.90	2210.72	1748.54	100.00
External causes	2780.39	2883.83	2696.75	2048.36	3449.62	1056.65
Unknown	1165.19	1165.79	1164.93	1193.84	1141.52	
No condition	938.04	1016.79	907.99	871.23	948.81	

NOTES: The birth episodes for newborns are excluded.

Table 5. National Estimates of Charges for Hospital Care National Medical Care Utilization and Expenditure Survey, 1980.

	Estimated Number of Discharges in thousands	Estimated Total Charges in millions	
		Survey	Real
Total	**$35,700**	**$79,961**	**$78,224**
Malignant neoplasms	1,540	7,583	8,606
Nervous/Sensory	1,943	2,886	3,013
Circulatory	4,747	12,780	12,472
Respiratory	3,356	6,736	7,982
Digestive	3,480	8,187	8,413
Urinary	3,344	7,001	5,312
Pregnancy	2,272	2,913	2,950
Musculoskeletal	2,133	5,346	5,353
External causes	3,856	10,721	10,398

NOTES: The birth episodes for newborns are excluded.

Table 6. Hospital Charges by Control and Imputation Status. National Medical Care Utilization and Expenditure Survey, 1980

	Imputation Status		
	Total	Imputed	Real
	Estimated Mean Charges		
Total	2239.78	2321.61	2191.12
Private			
Not-for-profit	2225.62	1955.47	2365.36
For-profit	2113.09	1892.47	2194.58
Government			
Non-federal	2188.97	2711.74	1824.52
Federal	3419.64	4607.82	6.43
Unknown	1782.61	2038.73	1523.28
		Number in Sample	
Total	2710	1001	1709
Private			
Not-for-profit	1816	617	1199
For-profit	200	52	148
Government			
Non-federal	447	182	265
Federal	114	87	27
Unknown	133	63	70