AN EVALUATION OF WEIGHTED HOT DECK IMPUTATION FOR UNREPORTED HEALTH CARE VISITS

Brenda G. Cox and Ralph E. Folsom, Research Triangle Institute

The National Medical Care Expenditure Survey (NMCES), a panel survey, was sponsored by the United States Public Health Service under funding from the National Center for Health Services Research. The purpose of this survey is to provide detailed information on the health of the residents of the United States, how and where they receive health care, the cost of the services, and how these costs were paid. The reported results from this study will ultimately have an impact on public policy concerning health care for the entire nation.

Much of this data could only be obtained in a household interview. For this reason, NMCES selected a stratified cluster sample of 13,500 households to represent the civilian, noninstitutionalized residents of the United States and interviewed them during the 1977 calendar year (Cox, Piper, and Frankel, 1980). Since estimates of health care utilization and expenditures obtained from household-reported data are known to be subject to bias, NMCES also included a record check survey component in which a one-third subsample of the NMCES sample individuals had their medical providers surveyed as a part of the Medical Provider Survey (Cox and Folsom, 1980). For individuals included in the Medical Provider Survey (MPS) with responding medical care providers, two data sets are available concerning the medical care visits made by the sample individual. The first set is the household-reported version of the visits made to the provider with the associated charges and diagnoses. The second set is the provider-reported version of the visits made by the respondent, again with the associated charges and diagnoses. Since the medical care provider had access to the respondent's medical records, it would seem reasonable to assume that the medical care data reported by the provider are more accurate than the household-reported medical care data. However, provider data are available for only 10,611 sample individuals whereas household data are available for 38,815 sample individuals.

This is the classic situation in which the data analyst can use double sampling regression estimation to increase the precision of survey estimates. These double sampling estimators would correct the estimate obtained from household data based upon the MPS-observed relationship between the household report and the provider report. In practice, double sampling estimation can be a time consuming process because of the need for model building investigations of the relationships between the two sets of data. Since the analysis plan for NMCES required the production of a very large quantity of tabular summaries, the only solution immediately available was to produce analyses based strictly upon the MPS-subsampled individuals. But this was itself undesirable because it required that the data for the 28,204 non-MPS individuals be ignored.

To avoid this problem, it was decided to investigate the potential for imputing data, as the provider would have reported it, to all NMCES sample individuals with this data missing. Further, the specification was made that estimates obtained using the provider-reported data, whether imputed or real, were to be equal in expectation over repeated imputations to the conventional double sampling estimate. Three tasks were implemented to investigate the potential for this type of provider data imputation. The first task was a comparison of household and provider reports of charges for the medical care visits that both reported (Williams and Folsom, 1981). The second task was to compare household and provider reports concerning the diagnosis associated with the medical care visits made by the household (Cox and McGrath, 1981). The third and final task dealt with unreported medical care visits. Household underreporting leads in the MPS sample to the phenomena of "discovered visits"; that is, the providers will report visits for which no household data exist. This paper documents the results of an evaluation of the use of weighted sequential hot deck imputation of discovered visits to individuals who were not selected for inclusion in the Medical Provider Survey.

1. The Matching of Household and Provider Data

The first step in the imputation of the provider-reported health care data involved the matching of household and medical provider utilization data. Matching the medical provider data set and the household data set resulted in reported visits of three types:

1. Matched visits. These are those visits reported in the household interview which were judged to have been matched to visits reported by the respondents' medical providers.

2. Unmatched visits reported by the household. These are the visits which were reported in the household interview to which no medical provider reported visits could be matched under the rules for matching.

3. Unmatched visits reported by medical providers. These visits included all visits reported by medical providers which could not be matched to household data.

The unmatched visits are either visits actually reported in one survey but not in the other or they are visits reported in both surveys which could not be matched using the selected matching rules (Cooley and Cox, 1981). In reviewing the results of this investigation of unreported health care visits, the reader should bear in mind that the accuracy of the matching directly affects the quality of the statistics presented in this report.

2. Description of the Imputation Procedure

There was no straightforward means of developing an imputation equation based upon the household survey data which could ascribe

unreported visits and the associated data records to non-MPS sample individuals or to MPS sample individuals with nonresponding medical providers. For this reason, utilization data for discovered visits detected in the Medical Provider Survey had to be imputed directly as unreported visit data to non-MPS sample individuals and to MPS sample individuals with nonresponding medical providers.

By adapting a sequential sample selection method discussed by Chromy (1979), a weighted sequential hot deck procedure was developed for use in imputing discovered visits (Cox, 1980). The procedure was designed so that means derived from the inputation-revised data would be equal in expectation to the weighted mean or proportion estimated using discovered visit data obtained only as a part of the MPS. Noting that variances, covariances, correlations, regression coefficients, and other higher order population parameters based on the set of variates to be imputed are estimated by simple functions involving weighted means of squared terms and cross-products, the expectation of such higher order statistics over repeated imputations should also reproduce the corresponding MPS-based estimators. Furthermore, since subpopulation proportions are estimated as weighted means of zero-one indicator variables, the mean-preserving property of the weighted sequential hot deck procedure reproduces in expectation the entire distribution of donor data. This property was doubly important for NMCES imputations since the donors were a probability subsample of the full NMCES sample.

There are two classes of NMCES sample individuals with possible unreported visits not detected through the MPS. First, individuals who were not selected for inclusion in MPS did not have any of their providers interviewed and hence provider data were totally missing for these individuals. Second, individuals who were selected for inclusion in MPS had incomplete discovered visit data when one or more of their providers failed to respond. Because of this difference in the type of nonresponse encountered, discovered visit imputation was done separately for MPS sample individuals and non-MPS sample individuals.

For individuals selected for inclusion in MPS, imputation occurred at the person/provider level. The imputation procedure involved linking responding person/provider combinations to the nonresponding person/provider combinations using the weighted sequential hot deck imputation algorithm.

Those individuals who were not selected for MPS have totally missing provider data. For these individuals, the imputation occurred at the individual level. In this situation, the non-MPS sample individual was linked to a MPS sample individual, again using the weighted sequential hot deck approach.

3. Determination of Imputation Classes and Sorting Variables

The weighted sequential hot deck procedure was implemented within imputation classes so that within each class, the expected value over all imputations of the weighted mean of the imputed data, estimated using the NMCES household sample weights, would equal the weighted mean of the MPS data estimated using MPS weights. The procedure is similar in effect to weighting class adjustment for nonresponse and hence the rules for determining variables to use in forming imputation classes are similar to those one would use for a weight adjustment procedure. Within imputation classes, the data records were sorted to provide greater control over the imputation process. Variables which were considered as candidates for use in forming imputation classes and as sorting variables were the following: age, race, sex, income of family, education of the individual, education of the household head, insurance coverage, and the respondent's assessment of his health status. All of these variables relate to socio-demographic characteristics of the individuals or their families. In addition to these variables, household survey data were available to classify the individuals according to the number of ambulatory visits and the number of hospital visits that were reported, with person/provider combinations further differentiable by type of provider.

Separate estimates of the average number of discovered visits were made for providers to whom visits were reported by the household and for providers to whom no visits were reported by the household (i.e., providers that the household stated were their usual source of medical care but to whom no visits were made). With respect to defining imputation classes, the summary statistics revealed that variability existed for all of the variables examined and was significant for the reporting domains defined by age, race, sex, income, and insurance coverage. In selecting the variables to be used in classing, it was decided that sex should be included to prevent imputing visits from males to females or vice versa. Because of their critical importance as reporting domains, it was further decided that race and age were to be used to the fullest extent possible. The income variable showed important variation in discovered visits over its levels but contained a large proportion of individuals with missing data. For this reason, it was decided to use the education of household head variable in lieu of income. The variability in discovered visits found between the levels of the variables for number of household-reported ambulatory visits and number of household-reported hospital visits was important especially between zero visits versus some visits reported. For this reason, a none versus some household-reported visits indicator variable was created for use in classing. These five variables were all that could be used for classing and sorting due to sample size restrictions.

Although the levels of insurance coverage and health status exhibited strong variation in the distribution of discovered visits, they were not used in the imputation process. Insurance coverage showed the most variation in discovered visits between those with public insurance only versus the remainder. It was felt that classing by age, race, sex, and education of household

head would capture much of this variability due to the correlation between these variables and the insurance coverage variable. Much the same reasoning led to not using the perceived health status variable.

Within each of these imputation classes, the records were sorted by an education of household head variable and then by the age of the individual in years. Serpentine sorting was used so that for the first level of the education variable the records were sorted by age from low to high values, for the second education level sorting from high to low age values, and for the third education level from low to high ages again. By using serpentine sorting, records which were adjacent in the data file were as alike as possible with respect to the education of the household head and age.

4. Evaluation of the Imputation Strategy

Thirty-two percent of the NMCES sample individuals had their providers included in the Medical Provider Survey. This implies that 68 percent of the NMCES sample individuals will automatically have provider data missing and hence be subject to discovered visit imputation. Even though the imputation procedure would use a probability subsample of the NMCES sample individuals (i.e., the MPS sample) to impute data to the remainder, it was felt that imputation to such a large proportion of the sample required an empirical evaluation of the imputation bias and variance before implementing the procedure.

To reflect the double sampling nature of the imputation, a test data set was constructed using data from individuals with one or more providers participating in the Medical Provider Survey. The basis for the evaluation was the discovered visit data obtained as a part of the provider survey. An evaluation data set was constructed by subsampling the MPS data set in such a manner as to mimic the structure of the MPS sample versus the NMCES sample. For records in the evaluation data set, the provider data for the subsampled individuals were used to impute data to the remaining individuals. In actuality, six evaluation data sets were constructed by selecting three independent subsamples and using the data for each of the three subsamples to impute data to the remaining individuals in two separate runs. By independently replicating the subsampling and imputation steps, measures could be obtained for the variability induced in survey estimates by the MPS subsampling and imputation.

No attempt was made to evaluate the effect of missing responses from providers for individuals who had at least one of their providers responding. To simplify matters, the evaluation was conducted as though there had been no provider nonresponse. The missing person/provider level data were due to nonresponse rather than subsampling and hence its missingness could not be simulated as easily as that of the missing person-level data.

In order to evaluate the quality of the imputations, the imputation-revised evaluation data sets were used to compute summary statistics of interest. Specifically, the average expenditures per individual for discovered visits and the average number of discovered visits were estimated for the total domain and for domains defined by age, by race, and by sex. Finally, the proportion of individuals who fell into cells defined by number of discovered visits and by total expenditures for discovered visits was calculated. Using MPS sample weights, these estimates were computed for each of the six imputation-revised evaluation data sets and for the data set of original responses to the MPS. These statistics formed the basis for the analysis of the bias and variance induced by imputation.

Independent subsampling and replicated imputation within subsamples were used so that the variance due to subsampling, $\sigma^2(S)$, and the variance due to imputation, $\sigma^2(I)$, could be separately estimated. The significance of these two variance components is this. Suppose one makes estimates for the full NMCES sample using the imputation-revised discovered visit data. These imputation-based estimates will contain a variance component due to the subsampling used in selecting MPS and a variance component due to imputation. The effect of these components can be approximated by $\sigma^2(S) + \sigma^2(I)$. The additional variability induced by MPS subsampling and imputation is sometimes large in comparison to the size of the actual estimate. This variability would of necessity have an adverse effect on the quality of estimates derived from the imputation-revised data.

Since the MPS sample is fixed, the variability induced by the MPS subsampling cannot be reduced. However, discovered visits may be imputed multiple times and the imputation variability reduced. If imputation is independently performed n times and the estimates derived from the n imputation-revised variables combined, the effect of imputation is reduced to $\sigma^2(I)$ and hence the overall effect of subsampling and imputation to $\sigma^2(S) + \sigma^2(I)/n$. The proportion of the variance which can be reduced by combining over n imputations is

$$[\sigma^2(S) + \sigma^2(I)/n]/[\sigma^2(S) + \sigma^2(I)].$$

Note that when n is one, this quantity is automatically equal to one.

In general, there is little reduction in variability obtained by using multiple imputations; this reflects the fact that the sampling component of the variance is substantially greater than the imputation component of the variance in all but a few cases. It should be noted that all of the domains for which Table 1 gives estimates were used in the imputation as classing or sorting variables. Thus, much of the imputation variance may have been removed by the control imposed on these reporting domains.

The average of all the estimates resulting from the three subsampling operations and the two imputations per subsample is also given in Table 1. The standard deviation of the average may be obtained as

$$\sqrt{\sigma^2(S)/3 + \sigma^2(I)/6} \quad .$$

For comparison purposes, the "true" mean is given; this is the estimate obtained when MPS-derived data are used. The relative difference

Table 1. Estimates Derived from the Discovered Visit Imputation Evaluation

Statistic	σ_s^2	σ_I^2	Average Imputed Mean	"True" Mean	Relative Difference
Total Population					
Ave. Exp./PID	55.1836	14.1046	37.16	42.40	0.12
Ave. No./PID	0.0049	0.0001	1.15	1.13	0.02
Age: Less Than 17					
Ave. Exp./PID	34.2152	0.0194	17.42	26.90	0.33
Ave. No./PID	0.0155	0.0002	0.90	0.92	0.02
Age: 17-29					
Ave. Exp./PID	71.6814	0.3055	30.65	46.48	0.34
Ave. No./PID	0.0135	0.0006	1.19	1.21	0.02
Age: 30-54					
Ave. Exp./PID	1.0191	167.3939	35.17	37.14	0.05
Ave. No./PID	0.0036	0.0014	1.17	1.07	0.09
Age: 55-64					
Ave. Exp./PID	9.6795	9.4346	29.31	28.15	0.04
Ave. No./PID	0.0144	0.0050	1.30	1.30	0.00
Age: 65 and Up					
Ave. Exp./PID	4,510.9218	78.2527	112.29	101.29	0.11
Ave. No./PID	0.0235	0.0045	1.57	1.57	0.00
Race: White					
Ave. Exp./PID	74.7194	17.5680	37.23	42.83	0.13
Ave. No./PID	0.0046	0.0001	1.09	1.07	0.02
Race: Nonwhite					
Ave. Exp./PID	12.0482	3.5730	36.64	39.43	0.07
Ave. No./PID	0.0196	0.0005	1.58	1.54	0.03
Sex: Male					
Ave. Exp./PID	10.0132	65.4615	30.7533	31.03	0.01
Ave. No./PID	0.0090	0.0002	1.0350	1.01	0.02
Sex: Female					
Ave. Exp./PID	114.8346	3.6255	42.5583	51.99	0.18
Ave. No./PID	0.0031	0.0001	1.2500	1.24	0.01
Proportion of Individuals Arrayed by Number of Discovered Visits					
0 Visits	0.00012425	0.00000162	0.59	0.59	0.00
1 Visit	0.00001302	0.00000201	0.19	0.19	0.01
2 Visits	0.00006564	0.00000113	0.09	0.09	0.00
3-5 Visits	0.00002387	0.00000259	0.09	0.09	0.00
6 or More Visits	0.00003096	0.00000028	0.04	0.04	0.03
Proportion of Individuals Arrayed by Discovered Visit Expenditures					
0 Dollars	0.00006707	0.00000020	0.69	0.68	0.01
01-25.00 Dollars	0.00008050	0.00000068	0.15	0.16	0.03
5.01-50.00 Dollars	0.00001803	0.00000018	0.07	0.07	0.04
50.01-100.00 Dollars	0.00000763	0.00000005	0.04	0.05	0.05
100.01-200.00 Dollars	0.00000729	0.00000039	0.03	0.03	0.02
200.01 or More Dollars	0.00000206	0.00000020	0.02	0.02	0.05

Ave. Exp./PID represents the average expenditures per individual for discovered visits.
Ave. No./PID represents the average number of discovered visits per individual.

between the average imputed value and the MPS "true" value is also given (this is the difference between the two estimates divided by the "true" value). Note that for estimates involving distributions of individuals by number of discovered visits or by their expenditures for these visits, the relative differences between the two estimates are relatively low, in every case less than ten percent and most often less than five percent. For expenditures for discovered visits, the relative difference between the two estimates is much larger. For the total population, the relative difference for the expenditure estimate is 12 percent; over all domains the relative differences range from 0.89 percent to 34 percent. The highest values for the relative differences were found for domains defined by age. This may result from the fact that age could only be partially used for the creation of imputation classes.

To assess whether the differences between the average imputed values and the MPS values were significant, t test statistics were computed. These t statistics are given for each of the estimates presented in Table 1. At the five percent level of significance, the critical values for the t statistic are ±4.303. An examination of Table 1 reveals that none of the t statistics are significantly different from zero. This implies that the combined subsampling/imputation process is reproducing, in expectation, the MPS estimates. This is not a particularly surprising result since these reporting domains were incorporated into the imputation process as classing or sorting variables and hence the weighted sequential hot deck procedure controlled for variation due to the variables used to define these domains.

5. Results of the Empirical Investigation

The results of the evaluation of the effect of the imputations on survey estimates indicated that discovered visit imputation was sufficiently accurate as far as reproducing the distribution of the provider data for domains defined by race, sex, and age. However, the evaluation did not examine estimates for reporting domains not used in the imputation process. Therefore, the unbiasedness of imputation-derived estimates for reporting domains not used as classing or sorting variables is not known. Further, the variability added by the subsampling and imputation was large in comparison with the size of the parameter being estimated.

These results agree with those of Williams and Folsom (1981) concerning provider charge imputation. In addition, Cox and McGrath (1981) concluded that the relationship between the provider and household diagnostic reports was not strong enough to develop an imputation scheme for non-MPS sample individuals, without aggregating diagnoses into categories so general as to be analytically meaningless. These studies suggest that imputation of data to non-MPS individuals is not a feasible alternative to double sampling regression estimation. Basically, this is due to the fact that data must be imputed for an extremely large proportion of the NMCES sample and the fact that the household reports and provider reports are not correlated to the strong level required by this level of imputation.

However, these studies do suggest a strategy which NMCES could adopt as an inexpensive alternative to double sampling regression estimation. This alternative approach would adjust the MPS sample weights within weighting classes so that they reflect the weighted distribution of the full NMCES sample. This approach will yield stratum estimates equal to those expected over repeated imputations using the weighted hot deck imputation procedure. Domains of interest can be explicitly used to define weighting class variables in order to maximize the double sampling efficiency gain for reporting groups. When domains of interest cannot be used to define weighting classes, the reweighting approach used with MPS data may produce less biased estimates than those derived from the imputation-revised data for all NMCES sample individuals, since only data from domain members are incorporated in the estimates. This reweighting approach also eliminates the variability induced by the imputation procedure as well. Folsom (1981) demonstrates that the reweighting strategy is equivalent to a double sampling regression estimation method based upon a regression model which specifies a separate mean parameter for each weight class (i.e., a cell mean model). When such models incorporate categorized versions of the most important household survey predictors, they can compare favorably to the best fitting model incorporating the associated uncategorized variables.

The reweighting approach is limited by the number of weighting classes which can be defined. This is important when one considers that over and under reporting of health care visits and differences in household and provider reports of charges and diagnoses must all be considered in defining the classes. If a high correlaton had existed between household and provider data, the original MPS imputation approach could have been superior to this reweighting approach. Such correlation was not found and hence the best approach for NMCES would appear to be the reweighting strategy.

It is important to note that even when a reweighting strategy is used, some imputations will be required. To prevent bias in the survey estimates, missing provider data for MPS sample individuals must be replaced. This missing provider data resulted from provider nonresponse and from failure of sample individuals to sign permission forms authorizing the interview. With respect to imputation of missing provider data, two options are available. First, one could use present MPS imputations to the fullest extent possible and augment these as necessary. Second, one could ignore previous MPS imputations and use person/provider level imputation to replace the missing provider data in partial records. If the household-reported data for the replacement person/provider were used instead of the original respondent's data, the relationship between household and provider data would also be preserved.

6. Concluding Remarks

The most striking finding of this investigation was the extent of the discovered visit phenomenon. An average of 1.13 discovered visits per individual was estimated using MPS

data (see Table 1). The average expenditures per individual for discovered visits was estimated as $42.40. It is obvious that discovered visits must be accounted for in order to avoid serious bias in expenditure and utilization estimates.

The study results discussed in this paper are no more than an initial attempt to deal with the phenomenon of discovered visits. Further research needs to be done using the rich resources of the NMCES/MPS data base. The estimates obtained as a part of this study suggest that the level of household under reporting in health care surveys may be so large as to make interpretation of the the household data difficult. Certainly, estimating the impact of federal legislation on health care costs requires that much more be known about the nature and extent of household under reporting of health care events.

7. Acknowledgements

The authors would like to acknowledge the advice and assistance received from the Intramural Research staff of the National Center for Health Services Research (NCHSR). Particular appreciation is expressed to Dr. Daniel Walden and Dr. Gail Wilensky who provided valuable insight into the analysis issues which directly affected this investigation of under reporting of health care visits by households.

The research discussed in this paper was performed for NCHSR under Contract No. HRA-230-76-0268. The views expressed in this paper are those of the authors and no official endorsement by NCHSR is intended or should be inferred.

8. References

Chromy, James R. (1979). Sequential sample selection methods. Proceedings of the American Statistical Association, Survey Research Methods Section.

Cooley, Philip C. and Brenda G. Cox (1981). A procedure for assessing errors in matching medical provider data with household data. Proceedings of the American Statistical Association, Survey Research Methods Section.

Cox, Brenda G. (1980). The weighted hot deck imputation procedure. Proceedings of the American Statistical Association, Survey Research Methods Section.

Cox, Brenda G. and Ralph E. Folsom (1980). The Sample Design and Weighting Plan for the Medical Provider Survey. RTI Report No. RTI/1320-13F. Prepared for the National Center for Health Services Research under Contract No. HRA-230-76-0268.

Cox, Brenda G., Lanny L. Piper, and Martin R. Frankel (1980). Development of Sample Weights for the National Medical Care Expenditure Survey. RTI Report No. RTI/1320-01F. Prepared for the National Center for Health Services Research under Contract No. HRA-230-76-0268.

Cox, Brenda G. and Debra S. McGrath (1981). The Relationship Between Household and Provider Reported Diagnoses for Visits Reported in the National Medical Care Expenditure Survey. RTI Report No. RTI/1320-20F. Prepared for the National Center for Health Services Research under Contract No. HRA-230-76-0268.

Folsom, Ralph E. (1981). The equivalence of certain double sampling regression estimators, weights adjustments, and randomized hot deck imputations. Proceedings of the American Statistical Association, Survey Research Methods Section.

Williams, Rick L. and Ralph E. Folsom (1981). Weighted hot-deck imputation of medical expenditures based on a record check subsample. Proceedings of the American Statistical Association, Survey Research Methods Section.