A MULTIPLE IMPUTATION APPROACH TO MICROSIMULATION

B. K. Atrostic, Congressional Budget Office
Ford House Office Building, Washington, D. C. 20515

KEYWORDS: Microsimulation, multiple imputation

Developing a microsimulation model of health care financing presents both conceptual and practical measurement problems. These problems arise for several reasons. The necessary data often are not collected jointly in a single survey or set of administrative records. Key economic relationships do not appear directly in the underlying data (such as the ultimate economic incidence of specific taxes, if that differs from the statutory or nominal payor). Estimates of behavioral responses to proposed changes are less than certain by nature, and depend on empirical estimates in the relevant literatures (such as response of the demand for health insurance to its price, or the response of health care expenditures to changes in copayment and coinsurance rates). Available estimates often apply to similar but not identical policy changes, were estimated for a somewhat different population group, or were estimated on data sufficiently old that behavior may have changed. Different modelers can make valid but different choices to resolve these problems. Each choice will produce different estimates of health care financing.

These problems suggest applying multiple imputation as an approach (rather than a specific technique or algorithm). Multiple imputation of these missing data, economic relationships, and behavioral responses (corresponding to alternative assumptions or empirical studies) can provide insights about the robustness of estimated policy outcomes. An explicitly multiple imputation approach can also provide a framework for consistent, rather than *ad hoc*, choices. Multiple imputation has been proposed in the statistical literature as a method of dealing with item nonresponse (e.g., Rubin 1987, and Little and Rubin 1987) that has better statistical properties than commonly used alternatives such as hot-decking, bootstrapping, or jack-knifing. It has been applied, in somewhat different ways, in current surveys (e.g., Kennickell 1994, and Shafer, Khare, Little, and Rubin 1993). Multiple imputation has recently been debated in the statistical literature (e.g., Efron 1994a,b, and Rubin 1994).

THE CBO MODEL: OVERVIEW

The CBO model is described in Atrostic and Bilheimer 1993. The model focuses on the receipt and financing of current health care services of the noninstitutional population. The CBO model will provide estimates of the sources of health care financing, the proportion of that financing that flows back to its sources as health care services, and whether those flows are the same in the aggregate as for population subgroups (such as income quintiles).

Estimating these flows requires estimates of actual spending for health care services on behalf of people in different income groups under the current system and under proposed changes. They also require estimates of the payments people in the same groups make to support the health care system--both directly through out-of-pocket payments and premiums, and indirectly through income and payroll taxes--under the current system and under proposed changes.

The CBO model follows the general microsimulation framework described in Citro and Hanushek, 1991. It is based on the Current Population Survey (CPS), supplemented by imputations of health expenditures, health insurance premiums, and health status from the 1987 National Medical Expenditure Survey (NMES), inflating to 1989 (or current) levels where necessary. Alternative information on premiums for employer-sponsored insurance comes from the 1989 Health Insurance Association of America (HIAA) survey of employers. The database is adjusted for the economic incidence of expenditures and financing (such as who actually pays the employer share of payroll taxes). Estimates of state and local taxes, and also the portions of federal, state, and local taxes (and the federal deficit) that finance health care spending all are added. The model is calibrated to reproduce external control totals, such as the National Health Accounts.

MULTIPLE CHOICES IN CONSTRUCTING THE PRIMARY DATA FILE

Constructing the primary data file for the CBO model presents two sets of opportunities for multiple choices: among models of health care expenditures, and among models of the joint distribution of health care expenditures and health insurance premiums. We calculate multiple alternatives for both choices. The premium and expenditure imputations discussed below have not been calibrated because it is easier to assess the basic similarities and differences between the imputations without them.

Modeling health care expenditures

A theoretical model of health care expenditures is implicit in each imputation of NMES health care expenditures to the CPS. Each imputation requires selecting specific variables from the CPS and NMES to match concepts in the theoretical model.

Expenditure model

In CBO's implicit model, health care expenditures depend on economic and demographic variables. This implicit model is consistent with theoretical and empirical models in the health literature. For its statistical matching, CBO models expenditures as:

(1) $E = f(U, I, D, S)$, where

E is total health care expenditures, U is the type of micro-level unit (individual, family, insurance unit, survey unit), I is income category (e.g., quintile of per capita or family income), D is a set of other primary economic and demographic categories (e.g., current insured status, source of insurance coverage, labor force status, age-sex category); and S is a set of secondary economic and demographic variables (e.g. years of age, dollar levels of income).

Modeling issues as reason for multiple imputations

The expenditure model takes this form in part because these variables appear on both the NMES and the CPS. This set of common variables is unlikely to include the full set of explanatory variables theory might suggest. For example, although self-reported health status and expenditures in the preceding period are commonly used to predict expenditures in the health and insurance literatures, they are not among the explanatory variables available for the expenditure imputation. Self-reported health status is collected on the NMES, but does not appear on the CPS. Expenditures in the previous period is not collected either in the NMES or the CPS. Other potentially important variables, such as measures of disability and labor force status, are defined differently enough in each survey that they are not comparable.

The form of this model is also conditioned by CBO's use of a statistical matching algorithm. Such algorithms require relatively broad categories within which NMES and CPS records must match exactly. (The use of finer categories results in a number of categories that rivals or exceeds the number of NMES observations.) The unit, income, and primary economic and demographic categories are the broad categories CBO uses. The algorithm also allows secondary categories within which records need not match exactly, but are linked based on their relative rankings within the broader categories. CBO uses year of age and actual income level.

The three alternative expenditure models CBO estimated differed in their micro-level units and secondary economic variables. Insurance units and family units were chosen because the two concepts are likely to differ. Insurance units are defined by current insurance industry practice. Spouses and dependent children typically can be covered under one policy, but other relatives in the household (such as adult siblings, or in-laws) typically can not. When survey households have such relatives, we create the appropriate number of insurance units and give each unit its own income and other demographic and economic characteristics.

The models also differed in the secondary economic and demographic characteristics included. In two models, we sorted the records in a match cell (where age group is part of the definition of a match) by the age of the primary person. Age is expected to be a determinant of health expenditures because health spending in the aggregate increases with age. In the third model, we sorted records in match cells (where income quintile is part of the definition of a match) by dollar levels of income as well as by age of the primary person. In this model, relative income levels would be expected to affect the level of health care spending, particularly if some spending is discretionary. In general, households with higher incomes spend higher total dollars on most broad categories of expenditures. Sorting by both age and income should increase the probability of matching units with comparable expenditures.

A model of health care expenditures and health insurance premiums

For analyses of health care financing a model of health expenditures alone may not be sufficient, if expenditures and premiums are jointly determined. And if they are jointly determined, they should also be imputed jointly to increase the likelihood of drawing correct inferences from the imputed data set. However, nonresponse in the premium portion of the NMES means that a joint imputation would not be straightforward.

Premium and expenditures model

The joint relationship of health care expenditures and premiums can be modeled several ways. Both variables may depend on the same explanatory variables (such as employment status, health status, and income), but not on each other:

(2a) $P = g(U, I, D, S)$ and
(2b) $E = f(U, I, D, S)$,

where U, I, D, and S are defined as in the expenditure model (equation (1)). That is, the variables explaining health care expenditures also explain the unit's health insurance premium (if any).

An alternative model is that premiums and expenditures, while having some explanatory variables in common, also depend on each other:

(3a) $P = j\ (U, I, D, S, E)$ and
(3b) $E = k\ (U, I, D, S, P)$.

That is, premiums depend on the unit's economic and demographic characteristics, but also on its health care spending; health expenditures depend on the unit's economic and demographic characteristics, but also on its health insurance premium (as a proxy for the unit's perceived cost of health care). More complex relationships between health care expenditures and premiums also could be specified.

Modeling issues as reason for multiple imputations

Any imputation of health insurance and premiums that applied models (2) or (3) (or a more complex variant of them) would require a data set in which all the explanatory and dependent variables were collected simultaneously for the same populations. However, the NMES survey structure, and its response rates, mean that such a data set is available only under strong assumptions about the nonresponse pattern.

Health expenditures were collected for the NMES household survey. Insurance premiums were collected from the households' insurers, for those households that gave NMES permission to contact their insurer. However, approximately 40 percent of persons with insurance refused to allow NMES to contact their employer or insurer to collect premium and plan specification data.

The simplest alternative is to assume that the nonresponse is random and ignorable. In practice, this means dropping from the NMES sample all premium nonresponders, and matching only premium responders to the CPS. Assuming random and ignorable nonresponse is what AHCPR has done for its NMES-based microsimulation model (Doyle et al. 1994). CBO can approximate this alternative by dropping observations for which an insured CPS respondent was matched to a NMES premium nonrespondent.

An alternative approach is to impute premiums separately. This is similar to applying only equation (2a), where the definitions of U, I, D, and S are changed to correspond to variables available in both the NMES premium responders sample and the CPS, and to rely less on detailed demographic characteristics. These revised variables include firm size and the employer share of employment-based premiums. Simple cross-tabulations typically show that premiums and employer-paid shares of premiums rise with employer size. For each imputed premium, there could be two alternative expenditure imputations: the expenditures imputed separately from the expenditure imputation (equation (1)) and the expenditures belonging to the NMES record whose premium was imputed.

Additional approaches are possible. The NMES could be "completed" by treating the missing premiums as a variable to be imputed within the NMES, and this completed NMES then imputed to the CPS. Premiums could also be imputed from an alternative data source, such as the most recent Health Insurance Association of America survey of employers. Although CBO has imputed the HIAA premiums, they are not discussed below because we have not had time to evaluate that imputation against other within-NMES alternatives.

EMPIRICAL DIFFERENCES ACROSS IMPUTATIONS

We find that the empirical importance of alternative imputations is difficult to predict. Some alternatives generate only small empirical differences in key relationships, while others generate much larger empirical differences.

Expenditure Imputation

The distribution of health expenditures is quite skewed, with roughly 10 percent of the population in the NMES survey accounting for 75 percent of total personal health care expenditures. Because the actual distribution is so skewed, comparing only means and variances between the NMES and the alternative imputations would not provide enough information to evaluate whether the imputations replicate the NMES distribution. Percentile distributions of total expenditures and three of its major components (out-of-pocket, private insurance, and Medicaid expenditures) provide additional insights. Percentile distributions for the subset of the population that actually has imputed expenses in each category are informative for similar reasons.

Univariate comparisons

Inferences about health spending vary surprisingly little across CBO's alternative imputations. The alternative imputations are relatively similar among themselves and similar to the NMES in terms of means, percentile distributions, and dollar levels of expenditure.

Dollar levels of spending for the population as a whole, estimated from NMES (in terms of means, standard deviations, and percentile distributions) are shown in the first column of Table 1. The three imputations (shown in the second, third, and fourth columns of Table 1) are virtually identical to the NMES for the mean and standard deviation, and for the 100th, 99th, 75th, and 50th percentiles. Out-of-pocket expenditures and private insurance expenditures also are distributed similarly in the NMES and the three imputations.

Table 1

	Sorted by:	Age	Age and Income	Age
	Match Varies by Unit Type	Insurance	Insurance	Family
	Actual NMES	**Imputed CPS**		
Mean	$1,481	$1,482	$1,474	$1,478
Standard Deviation	$5,147	$5,102	$5,060	$5,104
Percentiles:				
100th	$175,096	$175,096	$175,096	$175,096
99th	$22,535	$22,540	$22,512	$22,460
75th	$941	$939	$935	$939
50th	$274	$275	$274	$275
25th	$60	$64	$64	$64
5th	$0	$0	$0	$0

Because the distribution of health expenditures is skewed, we also reviewed the same statistics for those persons with expenditures in the relevant category. As for the full sample, distributions of total expenditures, out-of-pocket expenditures, and private insurance expenditures are similar to each other and to the NMES.

A few spending categories in some imputations appear to be quite different from each other and from NMES. For example, the Medicaid imputations for persons with any Medicaid expenses appear to differ substantially. However, the underlying NMES value for the 5th percentile of Medicaid spending is only $18 (in 1989 dollars); the implied value in the first imputation is 1.6 times the NMES value, but that is a small dollar difference, still $18.

Multivariate comparisons

The quality of the underlying statistical match can be checked in several ways. We find the matches to be close on all categories we examined. Although age and sex categories define imputation cells, and age is a sorting variable within imputation cells, the age categories are quite broad (roughly 15 years each). However, the actual age difference between a CPS record and its NMES expenditure donor averaged only 1.25 years in the preferred (insurance unit and age sort) imputation. In the same imputation, 79.9 percent of the CPS records were matched with NMES records in the same six broad categories (insurance type, sex, age, employment status, unit size, and income quintile), and 98.8 percent were matched in at least five categories.

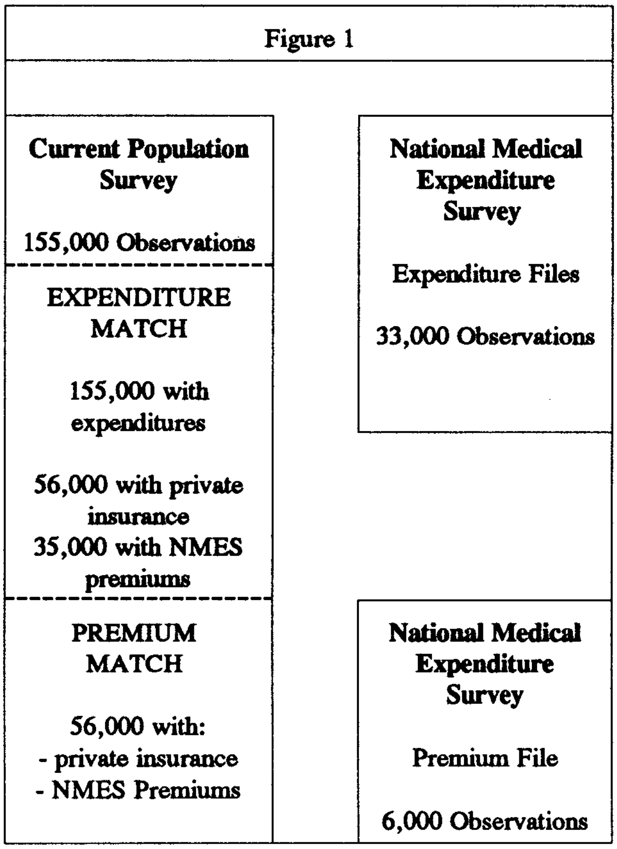
Premiums and Expenditures

We have examined two alternative imputations of NMES premiums and expenditures. For reasons of time and resources, neither imputation corresponds precisely to equations (2) or (3), because both are limited to the subset of records with employer-sponsored insurance. One imputation is the subset of the expenditure imputation records that come from NMES premium responders. Note that the expenditure imputation did not require that a privately insured CPS record be matched with a NMES premium responder. The expenditure imputation required only that a privately insured CPS record be matched with a NMES privately insured record. Using this subset as a "joint" imputation requires assuming that the premium nonresponse is ignorable.

A second imputation is the subset of the CPS with employer-sponsored insurance. Their premiums are imputed from NMES and their expenditures are those of the NMES record that provides their premiums. This subset allows us to look only at the relationship of premiums and expenditures for the privately insured. Expenditures for the uninsured and for those insured by government programs would have to be imputed separately. Both imputations are shown diagrammatically in Figure 1.

Because the distribution of premiums is relatively flat while the distribution of expenditures is highly skewed, we would expect the distribution of imputed premiums (given imputed expenditures) to differ from the distribution of imputed expenditures (given imputed premiums). It does. These differences, in turn, would imply different health care financing flows.

Figure 1


Current Population Survey
155,000 Observations
EXPENDITURE MATCH
155,000 with expenditures
56,000 with private insurance
35,000 with NMES premiums
PREMIUM MATCH
56,000 with:
- private insurance
- NMES Premiums
National Medical Expenditure Survey
Expenditure Files
33,000 Observations
National Medical Expenditure Survey
Premium File
6,000 Observations

Univariate

The distributions of total health expenditures (for persons with expenditures) differ between the imputation subsets. Their means differ by about $200, or more than 10 percent ($1825 for the premium imputation and $2023 for the expenditure imputation), and their skewness and kurtosis also differ. Their percentile distributions, however, are relatively similar. Their medians differ by about $12, or about 3 percent, and their interquartile ranges (the difference between the 75th and 25th percentiles) differ by about $100, or about 10 percent.

The means of imputed premiums differ by about $400, or more than 20 percent ($1723 for the premium imputation versus $2116 for the expenditure imputation), although their skewness and kurtosis measures are fairly similar. The percentile distributions of premiums from the two expenditure imputations also differ. Their medians differ by about $300, or about 15 percent. Their interquartile ranges (the difference between the 75th and 25th percentiles) differ by about $90, or about 6 percent.

Multivariates

The distribution of premiums within subgroups differs between the imputations. For example, total premiums for each income quintile have longer tails in the expenditure imputation. However, the interquartile ranges for each income quintile are similar in the two imputations. Similarly, the distribution of premiums and (log of) expenditures differ between the two imputations. Premiums are distributed more widely across expenditures in the premium imputation. By contrast, premium-expenditure combinations cluster at the lower range of premium values in the expenditure imputation. These distributions are shown in Figure 2.

CONCLUSION

Microsimulation models have inherent limitations. If microdata were available on all the relevant flows and cross-relationships, many elements of a baseline distribution could be estimated directly rather than developed through microsimulation. (Of course, important features of any baseline--such as allocating the employer share of payroll taxes--would not be given directly in any microdata source and would still have to be estimated.)

Each step in the imputation process may offer many alternative choices of concept and measure. Ultimately, the choice of an approach, or combination of approaches, offering a reasonable compromise in terms of plausibility, data requirements, and theoretical validity is a matter of professional judgement and experience. The methodology and assumptions underlying CBO's model will be updated as new data, research, and other information become available, and in response to comments and suggestions.

In order to assess the empirical importance of these choices, CBO makes a number of statistical matches using some of these alternatives. The multiple imputations provide the raw material for a series of sensitivity analyses. The alternative imputations provide the raw material for one purpose of the CBO model: demonstrating to users that estimates may be sensitive to theoretical and measurement decisions, that at times there will be more than one defensible set of choices, and that the complexity of the analysis should not be confused with precision and specificity. At the same time, however, the CBO analysis can be a useful guide to policymakers in understanding the qualitative story. Although the CBO model may not be able to describe the circumstances of individuals or narrowly defined groups, it can describe the circumstances of many groups of policy interest and determine whether the baseline distributions of services and financing vary among them.

The CBO microsimulation model is a team effort. The model was developed with Linda Bilheimer and Murray Ross, with programming and multiple imputations by Carol Frost. The model has benefited from comments by our colleagues at CBO. Assistance from Judy Shinogle in estimating the effects of alternative expenditure and premium imputations and in preparing graphics, and from Julia Jacobsen in compiling and preparing tables, is gratefully acknowledged.

REFERENCES

Atrostic, B. K. and Linda Bilheimer. 1993. "Modeling the Baseline Distribution of Health Services Spending and Payment," 1993 ASA Proceedings.

Citro, Constance and Eric A. Hanushek, eds. 1991. Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling. Volume I: Review and Recommendations. Washington, D. C.: National Academy Press.

Efron, Bradley. 1994a. "Missing Data, Imputation, and the Bootstrap," JASA, June 1994, 463-475.

Efron, Bradley. 1994b. "Rejoinder," JASA, June 1994. 478-479.

Farley, Dean, and Pat Doyle. 1994. "Alternative Strategies for Imputing Premiums and Predicting Expenditures Under Health Care Reform," this volume.

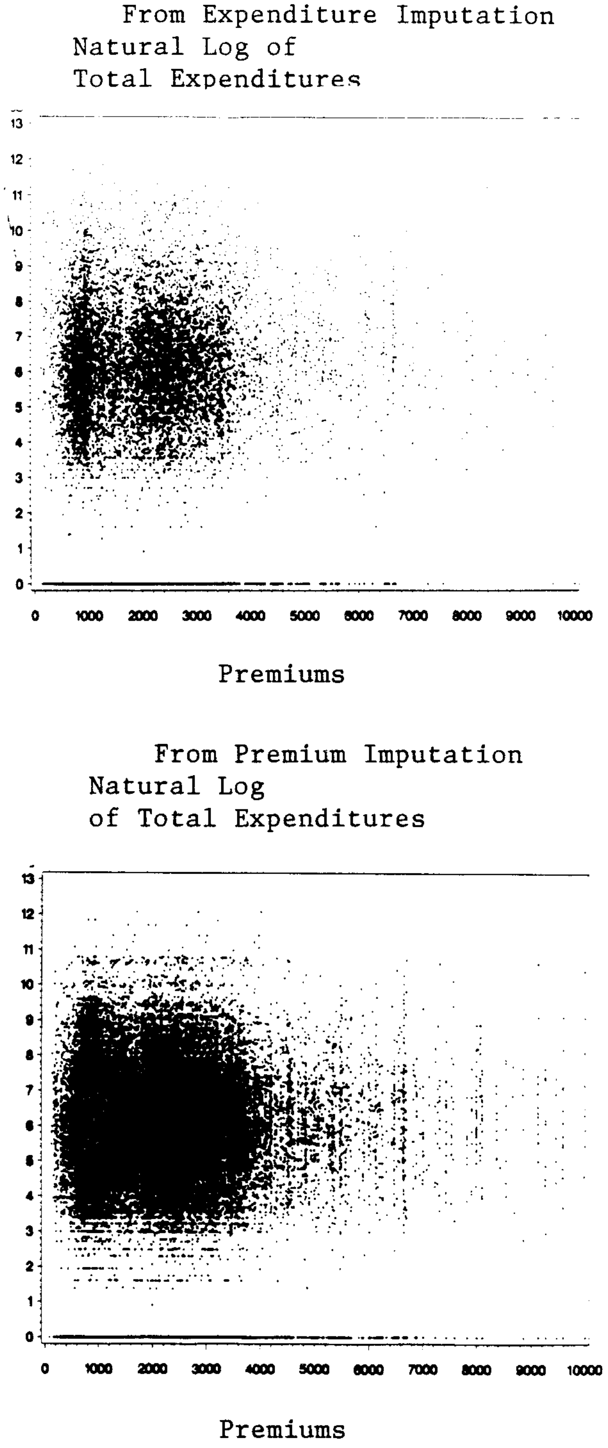
Kennickell, Arthur and Douglas McManus. 1994. "Multiple Imputation of Panel Data: The 1983-89 Survey of Consumer Finances," this volume.

Rubin, Donald B. 1994. "Comment," JASA, Jund 1994, 475-478.

Rubin, Donald B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Shafer, J., Khare, M., Little, R.A., and Rubin, D. 1993. "Multiple Imputation of NMANES III." Presented at the 1993 Annual Meetings of the American Statistical Association, San Francisco, CA.

Figure 2. Scatterplots of Log of Total Expenditures Versus Total Premiums



From Expenditure Imputation

From Premium Imputation