IMPUTATION OF MEDICAL COST AND PAYMENT DATA

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

Keywords: Gibbs Sampling, Hot Deck Imputation, Compositional Data

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

1. Introduction

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

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

2. Data Collection

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

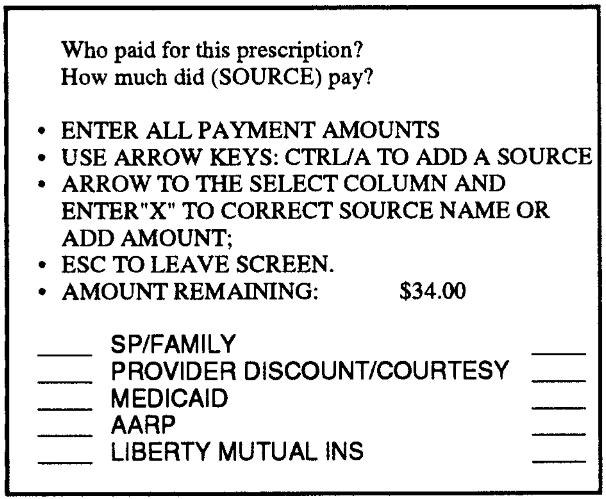
MCBS data are collected by CAPI (computer assisted personal interview). Interviewers carry laptop computers into the homes of Medicare beneficiaries and run a program that guides them through the interview. Figure 1 mimics a typical screen for collecting information about payments for a health care event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

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

3. Notation

Let $\delta=(\delta_1,\ldots,\delta_s)$ where $\delta_i=1$ if the i-th source is known to have made a payment, $\delta_i=0$ if the i-th component is known not to have made a payment. Given the structure of the interview, setting the delta's was not entirely straightforward. If there was an x

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

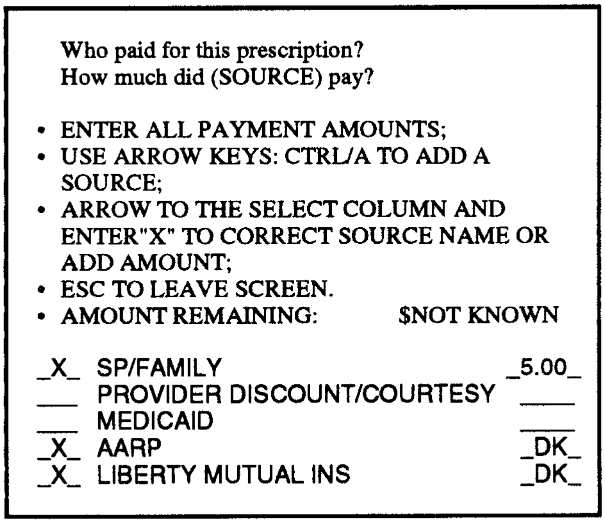


Who paid for this prescription?
How much did (SOURCE) pay?

- ENTER ALL PAYMENT AMOUNTS
- USE ARROW KEYS: CTRL/A TO ADD A SOURCE
- ARROW TO THE SELECT COLUMN AND ENTER"X" TO CORRECT SOURCE NAME OR ADD AMOUNT;
- ESC TO LEAVE SCREEN.
- AMOUNT REMAINING: $34.00

____	SP/FAMILY	____
____	PROVIDER DISCOUNT/COURTESY	____
____	MEDICAID	____
____	AARP	____
____	LIBERTY MUTUAL INS	____

Figure 1. CAPI screen prior to entering payment data

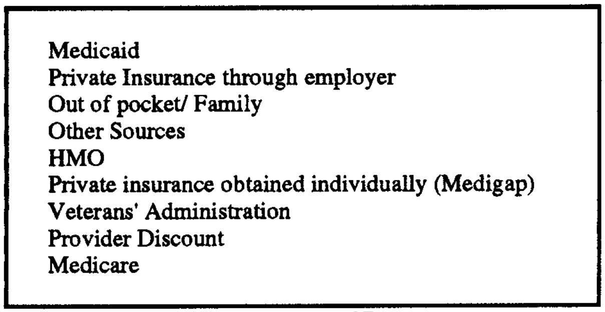


Who paid for this prescription?
How much did (SOURCE) pay?

- ENTER ALL PAYMENT AMOUNTS;
- USE ARROW KEYS: CTRL/A TO ADD A SOURCE;
- ARROW TO THE SELECT COLUMN AND ENTER"X" TO CORRECT SOURCE NAME OR ADD AMOUNT;
- ESC TO LEAVE SCREEN.
- AMOUNT REMAINING: $NOT KNOWN

_X_	SP/FAMILY	_5.00_
___	PROVIDER DISCOUNT/COURTESY	____
___	MEDICAID	____
_X_	AARP	_DK_
_X_	LIBERTY MUTUAL INS	_DK_

Figure 2. CAPI screen after entering partial payment data



Medicaid
Private Insurance through employer
Out of pocket/ Family
Other Sources
HMO
Private insurance obtained individually (Medigap)
Veterans' Administration
Provider Discount
Medicare

Figure 3. Sources of Payment

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

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

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

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

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

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

$0 \le Y_i \le Y_+$ for every i and (1)
$\Sigma_i Y_i = Y_+$. (2)

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

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

4. Data Editing and Exogenous Imputation

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

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

5. Missing Data Rates after Editing and Exogenous Imputation

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

To place these item nonresponse rates in context, although the rates are high compared to those typically experienced on surveys on other subject matters (such as labor force behavior), we do not view them as extraordinarily high for a consumer expenditure survey. People have a difficult time saving all receipts and bills for us over the typical four-month span between interviews. The few dollars spent as a co-payment for one container of medicine three months earlier do not constitute a very salient event in the typical respondent's memory. Furthermore, for those who are good about collecting receipts, many let them accumulate for months before submitting claims to insurance companies. Even with the longitudinal nature of the MCBS, it is difficult to track these claims over time. Most importantly, certain classes of beneficiaries have no knowledge of the cost of their prescription medicine; this is true for those who receive their drugs from the VA, from HMOs, through Medicaid, and through other public programs.

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

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

Table 1. Missing data rates

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

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

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

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

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

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

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

7. The Skeleton of the Algorithm

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

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

The second step is to impute δ. This is done slightly differently depending upon whether the total cost is known and whether there are any known payors with unknown amounts. However, the basic idea is the same: For each element h of Ω_h, conduct

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

a separate hot-deck run to impute the missing portion of δ, where the donors are chosen from among those cases that are already complete, the donors and missing cases are matched on X, the observed components of δ, and other available data. If the total cost is known, then that constitutes other available data that can be added to the match criteria (roughened into broad categories). If total cost is known and every known payor has a known amount, then the amount of money that must be covered by the missing deltas also constitutes other available data. Given the size of Ω_h and the three possibilities of reporting in Y and Y_+ for each element of Ω_h, a total number of 123 hot-decks were required for this step.[5]

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

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

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

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

8. Results

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

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

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

The covariance matrix of the delta vector, the covariance matrix of the Y vector, and the average payment amounts for each delta pattern were monitored as well throughout the imputation process. We noted that some correlations did change. It is

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

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

9. Limitations

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

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

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

10. Conclusions

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

References

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

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