Investigation of the Impact of Imputation on Variance Estimation in the Medical Expenditure Panel Survey (MEPS)[1]

Robert M. Baskin, John Sommers, and Trena M. Ezzati-Rice
Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850

Keywords: balanced repeated replication (BRR); Rao-Shao adjustment; Taylor method; collapsing; generalized variance function (GVF)

1. Introduction

The Medical Expenditure Panel Survey (MEPS) is a complex national probability sample survey sponsored by the Agency for Healthcare Research and Quality (AHRQ). MEPS, an ongoing annual survey, is designed to provide nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. MEPS consists of a family of three interrelated surveys with the Household Component (HC) as the core survey. The variable, healthcare expenditures, is considered one of the primary analysis variables in MEPS. The MEPS-HC, like most sample surveys, experiences item nonresponse despite efforts to collect complete information. There is a substantial amount of item nonresponse on expenditures in MEPS. To compensate for missing data, a weighted sequential hotdeck imputation approach is used to reduce the potential bias in estimating expenditures. The standard variance estimators do not account for any impact on variance due to imputation. The purpose of this study is to investigate methods for adjusting the variance estimates for variance due to imputation. Specifically, for imputing 2001 inpatient facility expenditures, two methods of adjusting the variance for imputation are studied and compared to the usual estimator of variance. Also, the coefficients of variation (CVs) from this study are compared with the CVs from the previous pilot study on outpatient expenditures, Baskin et al (2004). Finally an approach for providing users with the ability to adjust variances is discussed.

2. Background: MEPS Sample

The sample of households for the MEPS-HC is a subsample of households that responded to the prior year's National Health Interview Survey, conducted by the National Center for Health Statistics. The MEPS-HC uses an overlapping panel design in which data are collected through a series of five rounds of interviews over a two and one-half year period. Analytic weights, accounting for survey nonresponse, are calculated for MEPS and the details on the weights as well as detailed information on the MEPS sample design can be found in (Cohen, 1997; Cohen, 2000).

The variable, medical expenditures, is one of the primary analysis variables collected in the MEPS-HC. It is a composite variable, each component of which has some level of missingness. To compensate for the missing data and to improve the accuracy of the survey estimates, data on expenses for household respondents are also collected from a sample of their health care providers in the Medical Provider Component of MEPS. However, frequently expense data are not available from either survey. If expenditure data are missing, the missing values are imputed. The method used for imputation is the weighted sequential hotdeck described in Cox (1978). A further description of the methodology for imputing missing expenditure data can be found in Machlin and Dougherty (2004).

3. Impact of Imputation on Variance

The main advantage of imputation is that full case analysis can be employed on data with missing values and, if the imputation is effective, the correct multivariate structure of the data can be maintained. However, the use of standard software on data with imputed values assumes that all the data points are observed and doesn't take into account any variability in the imputation. Thus, standard variance estimates are downwardly biased. For a recent review of the state of the art of imputation, see Schafer and Graham (2000) and for a detailed source see Little and Rubin (2002).

The current method of imputation implemented by Westat for MEPS expenditure data is a form of weighted hotdeck. Although the survey weights are used to match donors to recipients, the hotdeck approach used includes a random number for selection of which donor to match to a recipient. This adds a level of variability to the data that is not reflected in the usual variance estimators.

4. Methods of Adjusting Variance for Imputation

The purpose of this study is to explore methods to estimate

[1] The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred.

the added variance due to imputation. Two types of methods have been previously published for estimating the impact of imputation on variance estimates. Multiple Imputation is a method of creating more than one imputation for each missing item and can be found in Little and Rubin (2002). There are also resampling methods to account for variance due to imputation. Two methods for bootstrap and two for jackknife, also mentioned in Little and Rubin (2002), are found in Rao (1996) and Fay (1996). For hotdeck imputation there is an adjustment for balanced repeated replication (BRR) variance estimators that can be found in Rao and Shao (1999). For the current imputation methodology in MEPS there are limitations with all of these approaches. For multiple imputation, our weighted hotdeck is not a proper imputation in the sense of Little and Rubin (2002). We could try to perform a proper imputation but the increased variance would be for the proper imputation and not the production imputation that we publish in the public use files (PUF). The bootstrap requires an imputation for each missing value in each bootstrap replicate and this is a priori too computationally intensive for this study. And finally, replicates are not created for public use MEPS data.

In the 2004 pilot study, Baskin et al (2004), we considered many methodologies but actually settled on a resampling method that, as far as the authors are aware, had not been previously implemented using a production imputation for a complex sample survey. For this follow up study, the same replication methodology was implemented with one modification. For the purpose of estimating the variance including increased variance due to imputation, we created 64 BRR replicates and independently reimputed missing data within each replicate as well as independently performing a full sample imputation. In the pilot study only 32 replicates were used to assess the impact of imputation of missing out-patient facility expenditures.

The imputation was carried out using the production software implemented by Westat. Because of the computationally intensive nature of performing the imputation 65 times, the imputation was only run for missing inpatient facility expenditures. The full sample as well as each set of 64 replicates was run through the inpatient imputation process in a manner mimicking the production process as closely as possible.

There are certain limitations to the results of the study, but together with the pilot it does provide important information about the impact of imputation on the variance of MEPS expenditure data. The use of 64 replicates in this study does satisfy the rule of thumb that at least 50 BRR replicates are considered the minimum required whereas the pilot study only used 32 replicates because of concerns about resources. In the 2001 PUF, inpatient facility events account for almost thirty percent of total expenditures.

Once the imputation runs were performed on the replicates, BRR weights were needed to calculate BRR estimates of variance which accounted for the increased variance due to imputation. These replicates and weights are not part of any PUF and only used internally for evaluation purposes, so they cannot be considered production quality.

5. Evaluation of the Methods

Once the full sample dataset along with the 64 replicated imputed datasets were available, the variances of estimates of mean, median and total inpatient facility expenditures were calculated using four methods for comparison purposes. For the purpose of this paper only the estimates and variances of total are reported. These variances of the total were calculated for the overall sample as well as for subsets of the sample corresponding to sex, race-ethnicity, education status, region and MSA status.

Two naïve methods of calculating a variance that ignored the imputation were used. First, using the stratification from the 2001 PUF, a standard SUDAAN weighted estimate of variance of total expenditures was calculated. Second, since BRR weights and replicates were available, a naïve BRR estimate of variance of total expenditures based on the full sample imputation only was calculated.

Two estimates of variance that account for imputation were also calculated. First, the BRR estimate of variance of total expenditures using the replicated imputation was calculated. There is also a method of adjusting the BRR variance estimate for imputation due to Rao and Shao (1999). This method requires the calculation of the full sample mean as well as a mean for each replicate within each imputation adjustment cell. The imputed data, but not the observed data, in each replicate are then adjusted by the difference in the full sample mean and the corresponding replicate mean. This method does not require reimputing the missing data in each replicate, but it does require knowledge of the cells used for imputation. Because of the complicated collapsing of imputation cells, the final set of imputation cells in each replicate as well as the final collapsed cells in the full sample is not known. This creates a difficulty in applying the Rao-Shao method but an approximation is still available and that was employed. The estimate can be applied to the uncollapsed cells and to cells that are as collapsed as much as possible and, on average, this gives lower and upper bounds on the adjustment. In the pilot study it was observed that collapsing of imputation cells increased the part of the variance *due to imputation*. This was again supported by the empirical evidence from the current study based on the two versions of the Rao-Shao adjustment.

6. Results

The estimate of standard error accounting for imputation using replication indicated that for the overall estimate of total inpatient facility expenditures, the increase was about 20% compared to the naïve SUDAAN estimates of standard error and the increase was about 30% compared to the naïve BRR estimates of standard error. The pilot study indicated an increase in the estimate of standard error accounting for imputation using replication for the overall estimate of total outpatient facility expenditures of about 30% compared to the naïve SUDAAN estimates of standard error. The increase in standard error estimate of 20% for the inpatient facility total was based on imputing about 28% of the items while the 30% increase in the outpatient facility total was based on imputing about 47% of the items. The following table, Table 1, gives the point estimates and standard errors of total inpatient facility expenditures for the overall sample as well as point estimates and standard errors of total outpatient facility expenditures for the overall sample from the pilot study.

The following graph shows the comparison of standard errors for inpatient facility expenditures computed by the four methods for the overall sample as well as all fifteen subgroups formed by the variables: sex, race-ethnicity, education status, region and MSA status. Note that the points are plotted with the x-axis corresponding to the sum of the weights associated with the subgroup.

Table 1.
Comparison of Inpatient and Outpatient Facility Events: 2001 (From research files, NOT official MEPS estimates)

All dollar estimates in billions

	Out-Patient Facility	In-Patient Facility
Events in sample	15,898	3,882
Estimated Population Events	149,459,732	34,965,227
Expenditures	$52.795	$230.035
% Events Imputed	47%	28%
SUDAAN SE	$2.538	$11.570
Naïve BRR SE	$2.567	$10.306
Imputed BRR SE	$3.313	$13.421
Rao-Shao Adjusted SE	$3.305	$12.871

(insert graph 1)

7. Discussion

Because replicate definitions, replicate weights, and replicate imputations are not available on the MEPS PUF, a user would not be able to account for the increase in variance due to imputation as was done in this study. In order to provide users with the information from this study, one option to consider is the feasibility of providing users with generalized variance functions (GVF) that account for variance due to imputation. Variances accounting for imputation could be calculated for many subgroups of the data and GVFs could be fit to data points which are pairs of sums of weights and the estimated variances. These GVFs are assumed to have a functional form of $\text{var} = a(weight)^2 + b(weight)$, i.e., the intercept term is assumed to be zero. Graph 1 above was a precursor of this idea and the general shape of the square root of the GVF would follow the curve for the plotted standard error. Using the fifteen subgroups cited previously, GVFs for all four types of standard errors were fitted using the lm (linear model) package in R, R

Development Core Team (2005). These GVFs are plotted in graph 2.

(insert graph 2).

The GVF curves appear to fit the points well, with the exception of some of the points corresponding to Region. The GVFs are easy to calculate once the variance estimates have been made, thus it appears that GVFs are an option to provide users with estimates of variances that account for imputation. Note that any method of estimating the variance accounting for imputation could be used in conjunction with the idea of the GVF. Thus multiple imputation, the Rao-Shao adjustment using BRR or replicating the imputation could be used and the results could be provided through a GVF. A possible limitation is that the GVFs would need to be produced for each event separately.

There is one further issue to be dealt with in terms of GVFs. For MEPS, a typical rule of thumb is that if a coefficient of variation (CV) of a standard error (the relative standard error) exceeds 30 percent the result should be flagged for publication purposes. For each of the methods of estimating a variance and for each of the subgroups, a CV of the point estimate was calculated. For subdomains with over 7.5 million estimated events the CVs were nearly always smaller than 10% but as the subdomain size decreased the CVs varied between 10% and 40% with the CV approaching 80% within the smallest subgroups. These results are shown in graph 3.

(insert graph 3)

For the GVF option, a discussion would need to be included on how to warn users of the GVFs on the size of the relative standard errors for small domain sizes.

A comparison of the CVs for the inpatient facility expenditure estimates and the outpatient facility estimates is given Table 2. Although there is clearly some variability the paired CVS are typically on the same order of magnitude.

Table 2.
Comparison of Inpatient and Outpatient Facility CVs: 2001

Group	In Patient cell size	Out Patient cell size	In Patient CVs	Out Patient CVs
Total	3882	15898	0.058	0.063
MSA	2787	4273	0.069	0.138
non-MSA	1095	11625	0.121	0.061
Region 4	1605	2771	0.101	0.193
Region 3	924	3812	0.146	0.155
Region 1	633	5064	0.147	0.118
Region 2	720	4251	0.147	0.125
Male	1457	6719	0.086	0.086
Female	2425	9179	0.061	0.06
Hispanic	587	1964	0.202	0.197
Black, not Hispanic	597	1921	0.135	0.196
Other, not Hispanic	2698	12013	0.065	0.066

8. Summary

This paper summarized a study of the feasibility of alternative methods to measure the impact of imputation on variance estimates for 2001 MEPS expenditure data. Specifically, estimates of variance for inpatient facility expenses from two replication methods (accounting for imputation) were compared to two naive estimates of variance. The empirical results from the current study demonstrate the viability of the replication approach and the results showed an increase in variance due to imputation for inpatient facility expenses. The results of this study were consistent with the earlier pilot study conducted in 2004 which studied the impact of imputation on estimates of variance for outpatient facility expenses. In this study, the increase in variance due to imputation for the overall estimate of total inpatient expenditures was about 20%. In comparison, in the pilot study, the increase in variance due to imputation for outpatient facility events was estimated to be 30%. The larger variance increase for outpatient facility events was associated with a 47% imputation rate compared to a 28% imputation rate for inpatient facility events. This study together with the pilot accounts for almost 40% of total expenditures based on the 2001 MEPS data.

More study is required before detailed advice and/or procedures can be provided that will enable users of MEPS data to account for variation due to imputation in analysis of MEPS expenditure data. With the goal of assessing the impact of imputation on total expenditures, methods need

to be explored for combining the results across event types. The replication methods used in this study should be compared with other imputation variance procedures, such as multiple imputation, and evaluated in terms of both results and ease of application to MEPS data.

References

Baskin, R.M., Wun, L., Sommers, J., Zodet, M., Machlin, S.R., Ezzati-Rice, T.M. and Saha, S. (2004) "Investigation of the Impact of Imputation on Variance Estimation in the Medical Expenditure Panel Survey", *Proceedings of the Survey Research Methods Section*, American Statistical Association, to appear.

Cohen S.B. (1997). "Sample Design of the Medical Expenditure Panel Survey Household Component". Agency for Health Care Policy and Research, MEPS Methodology Report, No. 2, Rockville, MD., AHCPR Pub. No. 97-0027.

Cohen S. B. (2000), "Sample Design of the 1997 Medical Expenditure Panel Survey Household Component". Agency for Healthcare Research and Quality, MEPS Methodology Report, No. 11, Rockville, MD., AHRQ Pub. No.01-0001.

Cox, Brenda (1980), "The Weighted Sequential Hot Deck Imputation Procedure", *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 271-276.

Fay, R. E.(1996), ``Alternative paradigms for the analysis of imputed survey data (Pkg: p473-520)", *Journal of the American Statistical Association*, **91** , 490-498

Little, R.J.A. and Rubin, D.B. (2002) Statistical Analysis With Missing Data (2nd Ed.). New York: Wiley.

Machlin SR and Dougherty D (2004), "Overview of Methodology for Imputing Missing Expenditure Data in MEPS", *Proceedings of the Survey Research Methods Section*, American Statistical Association, to appear.

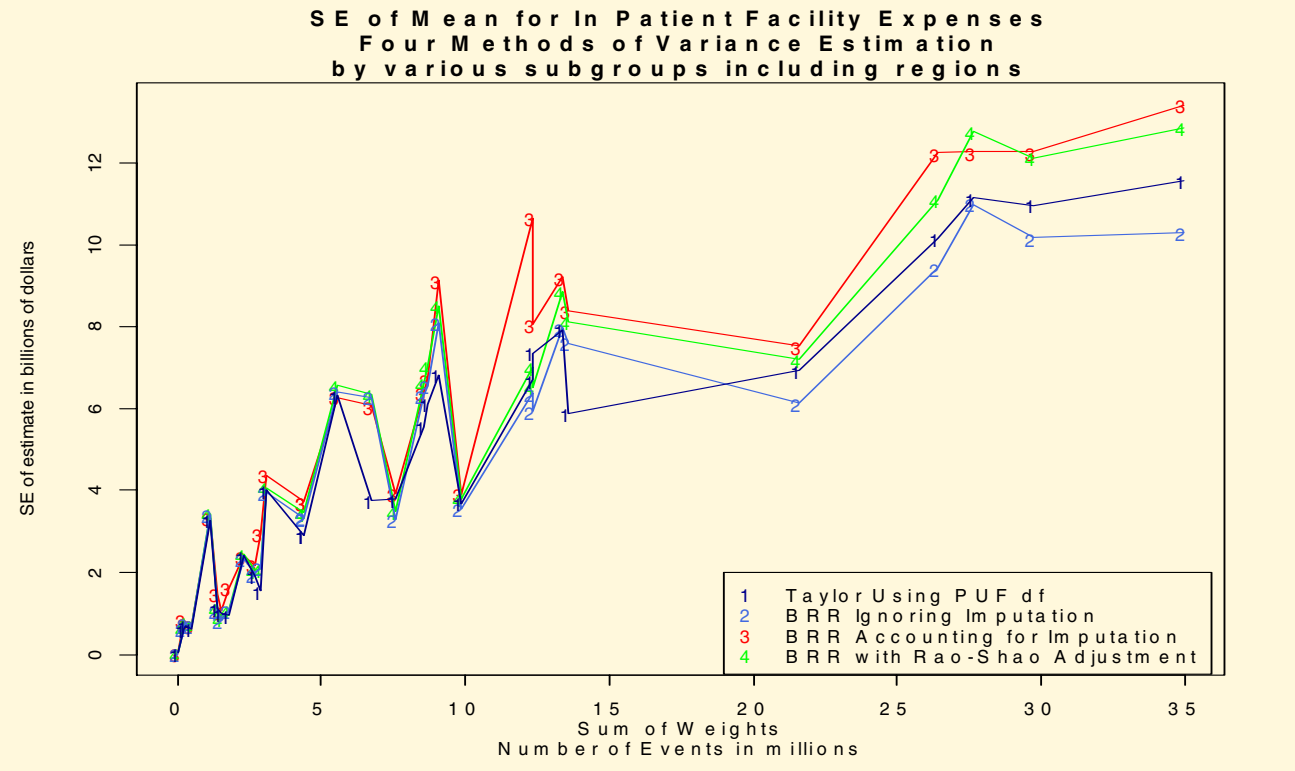
R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Rao, J.N.K (1996), ``On variance estimation with imputed survey data (Pkg: p473-520)", *Journal of the American Statistical Association*, **91** , 499-506

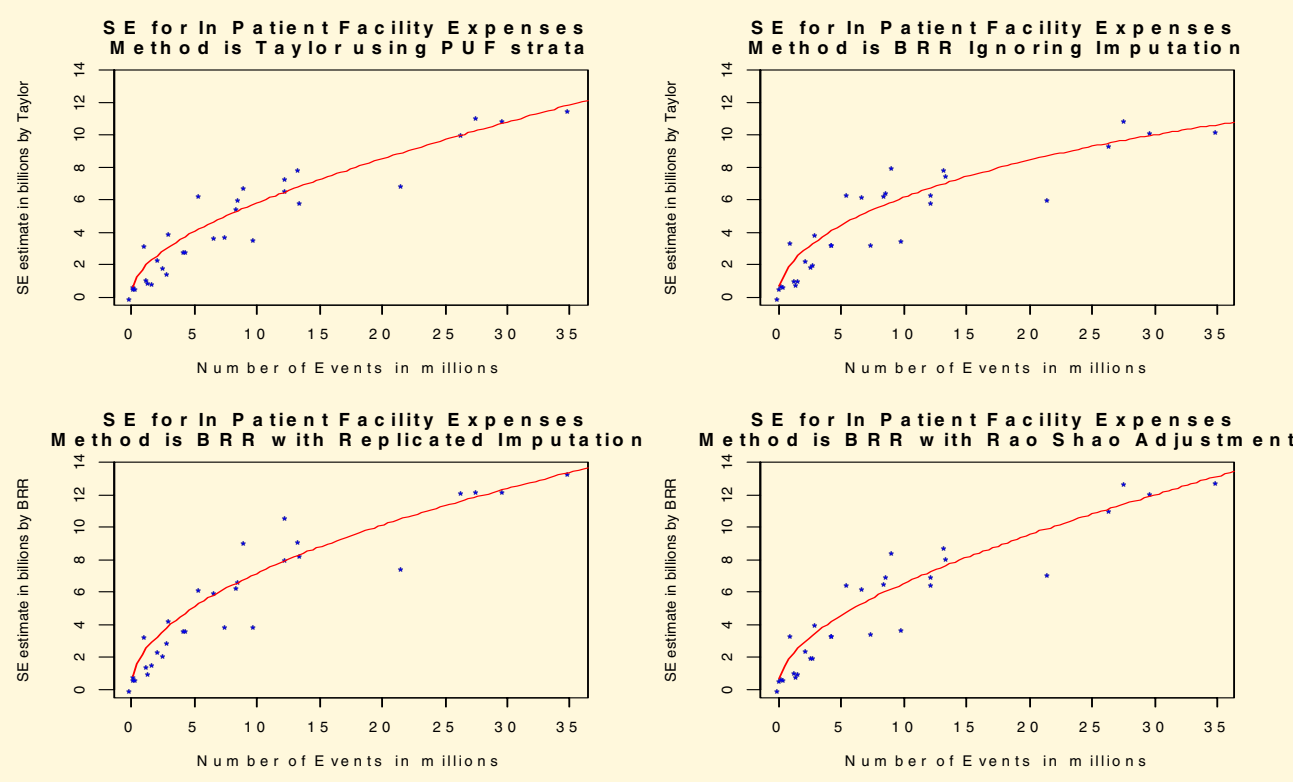
Rao, J.N.K, and Shao, Jun, (1999), "Modified balanced repeated replication for complex survey data", *Biometrika*, vol. 86 no. 2, pp 403-415.

Schaeffer, Joseph and Graham, John (2002), "Missing Data: Our View of the State of the Art", *Psychological Methods*, vol. 7 no. 2, 147-177.

Graph 1:



Graph 2



Graph 3