Comparison of Imputation Adjustment Techniques on Variance Estimation in the Medical Expenditure Panel Survey (MEPS)[1]

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

Abstract

The Medical Expenditure Panel Survey (MEPS) is a national probability sample survey designed to provide nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. Depending on the type of medical event, there are varying levels of item nonresponse on medical expenses as collected in the MEPS household interview. MEPS also collects expenditure data in the Medical Provider Component (MPC) of the survey. Missing expenditure data for health care events are completed through a weighted sequential hot deck procedure with MPC data as the primary donor source. Studies in 2004 and 2005 examined the impact of imputation on estimates of variance for MEPS health care expenditures. This study updates this research by investigating multiple imputation as a method to assess the impact of imputation on the variance estimates.

Key Words: Multiple imputation, variance estimation, Rao-Shao adjustment, survey data

1. Introduction

The Medical Expenditure Panel Survey (MEPS) collects data on health care utilization, expenditures, sources of payment, insurance coverage, and health care quality measures. The survey, conducted annually since 1996 by the Agency for Healthcare Research and Quality (AHRQ), is designed to produce national and regional estimates for the U.S. civilian noninstitutionalized population.

MEPS collects health care expenditure data from both household respondents (Household Component – HC) and from a sample of their health care providers (Medical Provider Component – MPC). Health care expense data are collected at the event level for each medical event type (e.g., office-based visits, hospital inpatient stays, etc.). While the amount of item nonresponse varies across the different medical event types, in general, there is substantial item nonresponse on the expenditure data in MEPS. When payment (i.e., expenditure) information is missing from either the household or medical provider components the missing data are imputed at the event level using a weighted sequential hot-deck procedure. The motivation for using an imputation procedure, such as the weighted sequential hotdeck, is to reduce the potential bias in estimating the expenditures that may be introduced if the missing data were to be ignored. However, when analyzing the imputed data, the standard variance estimators implemented in common statistical software packages do not account for any impact on the variance that is due to the imputation.

The purpose of this paper is to follow-up on two previous studies that investigated methods for adjusting variance estimates that account for the impact on variance due to imputation. Specifically, the objectives of this study are to evaluate the application of multiple imputation (MI) for deriving MEPS expenditure estimates and to compare the resulting variance estimates to those derived using the Rao-Shao variance adjustment (an adjustment technique considered in the previous studies).

[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.

2. Background

2.1 MEPS Overview

The MEPS-HC collects data from individual households and their members. These households come from a nationally representative subsample of households that participated in the prior year's National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics. The NHIS sample is a stratified, multistage, probability proportional to size (pps) selection of households. For most years, the sample comprises approximately 200 PSUs and covers approximately fifteen thousand households or about twenty-seven thousand persons. The approximate numbers of covered households/persons vary slightly from year to year. Using an overlapping panel design the MEPS-HC collects data over a two and one-half year period through a series of five rounds of interviews. Details regarding the MEPS sample design as well as the construction of analytic weights can be found in Cohen (1997 & 2000) and Ezzati-Rice (2008).

While the MEPS-HC collects data reflecting various facets of the U.S. health care system (e.g., utilization, insurance coverage, access to care, and quality) the primary intent of the survey is to collect data on health care expenditures. The survey facilitates analyses of distributions of health care expenditures and sources of payment, concentrations of expenditures, expenditures for specific health conditions, and trends in expenditures over time. The health care expenditure variables are the key analytic variables in MEPS and they are highly policy relevant.

It is difficult to obtain complete expenditure information from household respondents. In effort to reduce the level of item nonresponse for expenditures and to improve the accuracy of household reported data, the MEPS-MPC collects expenditure data from a sample of the respondents' health care providers. However, for a significant proportion of medical events, expenses are not available from either survey source (i.e., MEPS-HC or MEPS-MPC) and the data are imputed. These missing data are currently imputed using the weighted sequential hotdeck (Cox, 1980). A description of how the weighted sequential hotdeck is applied in context of the MEPS expenditure data can be found in Machlin and Dougherty (2004).

2.2 Previous Study Findings

Two previous studies have examined MEPS expenditure estimates and variance adjustment methods that account for the imputation's impact on the variance (Baskin et al. , 2004 & 2005). The impetus for these studies was the fact that most analyses of the MEPS expenditure data are performed using standard statistical software packages and assume that all the data values are observed. Treating all data values as observed does not reflect any potential variance introduced by the imputation procedure. Hence, the variance/standard error estimates reported from these analyses tend to be downwardly biased (i.e., too small).

To gauge the impact of the current weighted sequential hotdeck imputation on the variance estimates for MEPS expenditures, Baskin et al. compared two variance estimates that were derived ignoring imputation to two variance estimates that accounted for imputation. The two naïve variance estimates were derived using Taylor series expansion and balanced repeated replication (BRR). To generate variance estimates that accounted for the imputation the authors first used the BRR replicates and independently reimputed missing data within each replicate and for the full sample using the production software maintained by Westat. BRR estimates of variance were then generated which accounted for the added variance due to imputation. In addition, the authors used a modified BRR adjustment method developed by Rao and Shao. The Rao-Shao adjustment is performed at the replicate level and only imputed data values are adjusted. The adjustment made to the imputed values amounts to the difference between the full sample mean and the corresponding replicate mean. These studies looked at both MEPS inpatient and outpatient expenditure data from 2001. Their findings suggested an approximate 30% increase in the estimated standard error (SE) when accounting for the imputation.

3. Methods

The initial objective of this study was to evaluate the application of multiple imputation (MI) for generating MEPS expenditure estimates. Given that MI accounts for the impact on the variance due to the underlying imputation, it seemed intuitive to compare the findings of this paper to the above-mentioned studies. However, Donald Rubin suggests that when using MI it is highly desirable to choose an imputation method he refers to as *proper* (1987). A

proper imputation is one that allows MI to account for the uncertainty in estimating a given parameter; an *improper* imputation does not allow MI to fully account for any uncertainty in estimating a parameter (Rubin, 1987). Despite the fact that there is a small random component built into the weighted sequential hot deck currently used to impute MEPS expenditure data, the method is not considered *proper* according to Little and Rubin (2002)(also Rubin, 1987). Hence, it is ill suited for MI. The authors of the two previous MEPS studies acknowledged this and chose not to investigate MI. Given that MI is the desired means by which to generate variance estimates for this paper, it is first essential to choose a *proper* method by which to impute the missing data. That said, an additional objective of this research is to compare the variance estimates derived using MI to those derived using the Rao-Shao adjustment. The Rao-Shao adjustment was developed in the context of hotdeck imputation, so in order to compare the MI estimates to Rao-Shao estimates it is necessary to determine the applicability of Rao-Shao to the final chosen imputation method.

3.1 Data

Data for this project are the same hospital inpatient facility events from 2001 that were examined previously by Baskin et al. Hospital inpatient events are of particular interest for a number of reasons. First, these events represent a sizable proportion of overall health care expenditures: ≈29 percent. Second, inpatient expenditures are much more variable and more positively skewed than other types of medical event expenditures (e.g., office-based expenditures). Third, these data have a relatively large proportion of observations that require either full or partial imputation: ≈28 percent. Due to the resource intensive nature of creating an updated analytic file with current imputation classes, the decision was made to continue working with the 2001 data. Tabulations of more recent MEPS data suggest that the proportion of inpatient events requiring imputation has been consistent over the years.

3.2 Multiple Imputation

Single imputation involves replacing/filling in missing values with values based on information from the observed data. This can be achieved through either an explicit modeling approach, which uses a predictive distribution based on a formal statistical model (e.g., mean substitution, regression, stochastic regression), or an implicit modeling approach based on an algorithm, which implies an underlying model (e.g., hot/cold deck imputations, substitution, composite methods) (Rubin, 1987). Either approach results in a complete data set/column.

Multiple imputation is the act of performing an imputation d times (generally $d \geq 5$ times) resulting in D complete data sets/columns. Standard statistical analysis is then performed on each of the D complete data sets/ columns. The results of these analyses are then combined to produce MI inferences.

The MI estimate is the average of the estimates from each of the complete data sets/columns (Equation 1).

$$\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}_d, \qquad \hat{\theta}_d = \text{estimate from the } d\text{th completed data set/column}, \quad D = \#\text{ of imputed data sets/columns}$$

Equation 1: MI estimate (Little and Rubin, 2002); the estimate could be a total, mean, ratio, etc.

The MI variance is comprised of two components: the within-imputation variance and the between-imputation variance. The within-imputation variance is the average of the estimated variances for each D complete data set/column. The between-imputation variance is the variance of the estimates computed for each of D completed data sets/columns. Computational equations are presented in Equation 2.

within-imputation variance

$$\bar{W}_D = \frac{1}{D}\sum_{d=1}^{D} \hat{W}_d, \quad \hat{W}_d = \text{estimate of variance of } \hat{\theta}_d \text{ from the } i\text{th completed data set}$$

between-imputation variance

$$B_D = \frac{\sum (\hat{\theta}_d - \bar{\theta}_d)^2}{D-1}$$

Equation 2: MI variance components (Little and Rubin, 2002).

The total MI variance is the sum of the within-imputation component and the between-imputation component. There is a slight adjustment factor made to the between-imputation variance (Equation 3).

$$T_D = \overline{W}_D + \frac{D+1}{D} B_D, \quad \frac{D+1}{D} \text{ is an adjustment for a finite } D$$

Equation 3: MI total variance (Little and Rubin, 2002).

3.3 Statistical Software & Analysis

There are various statistical software packages/add-ons available on the market or as free-ware that are capable of performing MI analyses. Analyses for this project were performed using R; a statistical computing and graphics environment that is available for free at http://www.r-project.org/. There are two predominate ways to do MI in R: the `aregImpute` function in the `Hmisc` package authored by Frank Harrell and the `mice` function in the `mice` package authored by S. van Buuren and C.G.M. Oudshoorn. The analyses presented here were performed using `mice`, which stands for Multivariate Imputation by Chained Equations (MICE). This package provided three methods for imputing continuous/numeric variables such as the MEPS expenditures: unconditional mean imputation, Bayesian linear regression, and predictive mean matching (PMM). Unconditional mean imputation is not a *proper* method so it was not considered. Attempts to use Bayesian linear regression yielded negative imputed values, which were determined to be nonsensical in the context of health care expenses. Therefore, by process of elimination, PMM was chosen. PMM is a stochastic regression procedure which matches donors and recipients based on the proximity of their regression predicted scores (e.g., probability of missing expense data). The PMM implemented in `mice` uses Gibbs sampling to introduce uncertainty into the model parameters and is thus a *proper* imputation method.

Inpatient expenditure data were imputed independently for each of k=125 imputation cells using PMM (i.e., 125 PMM models were fit). Potential variables used as predictors included event and patient characteristics: length of stay, reason in hospital, age category, sex, MSA/non-MSA indicator, and Census region. Predictors were dropped from a model if they caused the model to be singular. The package defaults for the number of iterations (5) and imputations (5) were used. Once all the expenditure data were imputed survey estimates were generated for each column of complete data using the `survey` package authored by Thomas Lumley and MI inference equations were used to derive the MI estimates.

4. Results

Table 1 presents a sample case listing for one of the imputation cells. This table illustrates how the PMM resulted in five imputed/complete data columns. The predictors used in the PMM model for this particular imputation cell were length of stay, reason in hospital, age category, and sex. The original expenditure data with missing values are in the column labelled "Exp" and the resulting multiple imputations are labelled "PMM1" – "PMM5". The uncertainty introduced into the PMM model is evident across the five columns of imputed data.

Table 2 presents total hospital inpatient expenditures for each of the five complete data sets/columns as well as the estimated standard errors and variances of these estimates. These estimates ranged from $222 billion to $230 billion. The MI estimate of the total is simply the average of these estimates (approximately $226 billion). The MI estimate for the variance of this total is a very large number ($1.51E+20) derived by taking the average of the variances across the five complete data sets/columns and adding the (slightly adjusted) variance of the total estimates from each imputation. The square root of this variance estimate yields a MI estimate for the standard of the total of approximately $12 billion.

Table 1. Example case listing for an imputation cell (1 of 125 cells)

LOS	Reason	Age	Sex	Exp	PMM 1	PMM 2	PMM 3	PMM 4	PMM 5
1	1	1	1	2,388.10	2,388.10	2,388.10	2,388.10	2,388.10	2,388.10
3	2	1	1	Missing	1,792.43	890.44	489.91	1,792.43	2,155.41
2	2	1	1	2,155.41	2,155.41	2,155.41	2,155.41	2,155.41	2,155.41
11	2	1	1	Missing	34,620.35	34,620.35	34,620.35	489.91	34,620.35
3	2	1	1	Missing	1,792.43	890.44	489.91	1,792.43	2,155.41
2	1	1	1	Missing	2,931.17	535.54	535.54	2,388.10	2,931.17
9	2	1	1	Missing	34,620.35	34,620.35	34,620.35	489.91	34,620.35
4	2	1	1	1,792.43	1,792.43	1,792.43	1,792.43	1,792.43	1,792.43
3	2	1	1	Missing	1,792.43	890.44	489.91	1,792.43	1,792.43
3	3	2	2	Missing	2,931.17	2,931.17	2,931.17	2,931.17	2,931.17
1	1	1	1	2,134.91	2,134.91	2,134.91	2,134.91	2,134.91	2,134.91
4	3	2	2	Missing	2,155.41	535.54	489.91	489.91	890.44
3	3	2	2	2,931.17	2,931.17	2,931.17	2,931.17	2,931.17	2,931.17
1	1	1	1	0.00	0.00	0.00	0.00	0.00	0.00
9	2	2	2	10,067.20	10,067.20	10,067.20	10,067.20	10,067.20	10,067.20
2	5	1	2	489.91	489.91	489.91	489.91	489.91	489.91
3	4	1	2	890.44	890.44	890.44	890.44	890.44	890.44
3	3	1	2	535.54	535.54	535.54	535.54	535.54	535.54
3	2	1	1	Missing	1,792.43	890.44	489.91	1,792.43	1,792.43
8	1	1	1	Missing	34,620.35	34,620.35	10,067.20	2,155.41	10,067.20
7	4	1	1	34,620.35	34,620.35	34,620.35	34,620.35	34,620.35	34,620.35

MEPS hospital inpatient facility event data, 2001 (not official public release data)

Table 2. Total hospital inpatient expenditures

Imputation	Total (millions)	(SE) (millions)	Variance
1	$222,416	($11,460)	$1.31E+20
2	$223,683	($11,424)	$1.31E+20
3	$224,746	($11,521)	$1.33E+20
4	$226,449	($12,188)	$1.49E+20
5	$230,480	($12,333)	$1.52E+20

MEPS hospital inpatient facility event data, 2001 (not official public release data)

Similar results for mean hospital inpatient expenditures are presented in Table 3. Mean expenditures per inpatient event ranged from $6,361 to $6,592 across the five imputations. The MI estimates of mean expenditures per event and the corresponding variance and standard error are $6,451, $69,869, and $264, respectively.

Table 3. Mean hospital inpatient expenditures per event

Imputation	Mean	(SE)	Variance
1	$6,361	($239)	$56,911
2	$6,397	($231)	$53,543
3	$6,428	($238)	$56,778
4	$6,476	($261)	$68,119
5	$6,592	($257)	$66,065

MEPS hospital inpatient facility event data, 2001 (not official public release data)

The increase in variance when accounting for imputation is obtained by dividing the within-imputation variance into the total variance. For total hospital inpatient expenditures this is $1.51E+20 / $1.39E+20 ≈ 1.0843; an approximate

8% increase in variance due to imputation. For mean inpatient expenditures, the increase in variance due to imputation is approximately 16% (i.e., \$69,869 / \$60,283 ≈ 1.159).

Table 4 presents mean hospital inpatient expenditures by various population subdomains.

Table 4. Mean hospital inpatient expenditures per event, by population subdomains

Population		N	MI Mean	MI Variance	Total MI Variance / Within-Imputation Variance
Aggregate		3,882	\$6,451	\$69,870	1.159
Age	0-17	408	\$4,866	\$459,967	1.091
	18-44	1,155	\$4,166	\$50,117	1.044
	45-64	998	\$7,667	\$318,441	1.017
	65+	1,321	\$7,909	\$276,294	1.457
Sex	Male	1,457	\$8,087	\$232,595	1.013
	Female	2,425	\$5,438	\$71,643	1.351
Race/	White/Other, NH	2,698	\$6,641	\$106,751	1.229
Ethnicity	Black, NH	597	\$6,289	\$475,157	1.396
	Hispanic	587	\$4,893	\$139,658	1.111
Poverty	Poor/(-) Income	789	\$6,041	\$724,579	2.096[1]
Status	Near Poor	310	\$6,081	\$300,682	1.062
	Low Income	741	\$6,736	\$310,245	1.058
	Middle Income	1,125	\$5,977	\$159,176	1.174
	High Income	917	\$7,141	\$285,456	1.077

MEPS hospital inpatient facility event data, 2001 (not official public release data)
[1]unstable estimate

The increases in variance when accounting for the imputation varied greatly across subdomains. For example, the observed increase in variance due to imputation was approximately 35% for women's events, but only a modest 1% for men's events. Likewise, the observed increase in variance due to imputation is approximately 100% for poor people's events, but more modest for events experienced by higher income groups.

5. Conclusions

Overall, using MI, the observed increase in variance due to imputation is approximately 8% for the total inpatient expenditure estimate and approximately 16% for the mean inpatient expenditure estimate. The increases in variance due to imputation observed from the subpopulation analysis were highly variable. Some statisticians have been critical of MI's ability to provide valid inferences for subdomains, particularly in the context of complex survey data/large public use data. In particular, critics suggest that MI will often over-/understate the variances for subpopulations. The erratic nature of the increases observed in this study appears to support those critics of MI. These findings therefore suggest that applications of MI could be limited with the MEPS data except at the aggregate level. The usefulness of using MI to derive MEPS estimates is further limited by the fact that interest in MEPS data nearly always involves the derivation of estimates for subpopulations.

Application of the Rao-Shao adjustment method to the PMM imputation as part of this project was somewhat problematic. The challenge was in determining the PMM expectation at the replicate level in order to apply the Rao-Shao adjustment. This evaluation will require further collapsing of the imputation cells, thus future work will focus on an appropriate collapsing strategy.

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.

Baskin, R.M., Sommers, J., and Ezzati-Rice, T.M. (2005) "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.

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.

Ezzati-Rice, TM, Rohde, F, Greenblatt, J, *Sample Design of the Medical Expenditure Panel Survey Household Component*, 1998–2007. Methodology Report No. 22. March 2008. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr22/mr22.pdf.

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 SurveyResearch Methods Section*, American Statistical Association.

Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. New York: Wiley.