Imputation Variance Estimation by Multiple Imputation Method for the National Hospital Discharge Survey

Qiyuan Pan[1] and Iris Shimizu[2]
[1]Division of Health Care Statistics, National Center for Health Statistics, 3311 Toledo Road, Room 3427, Hyattsville, MD, 20782
[2]Office of Research and Methodology, National Center for Health Statistics, 3311 Toledo Road, Room 3123, Hyattsville, MD, 20782

Abstract[1]
The National Hospital Discharge Survey (NHDS) is currently conducted annually by the National Center for Health Statistics. This survey covers discharges from non-institutional, non-Federal, short-stay and general hospitals in the 50 States and the District of Columbia. Only three variables, namely age, sex and length of stay, are imputed for item non-response in NHDS data files. A hot deck method is used to impute the missing values. It is not known how this data imputation affects the variance approximations for estimates involving imputed values because the variance due to the imputation has not been evaluated for NHDS. This paper discusses an application of the multiple imputation method to estimate the magnitude of imputation variance for the 2006 NHDS.

Key Words: Health care survey, hot-deck imputation, complex sample survey

1. Introduction

1.1 Purpose

It is usually inevitable that data from a health survey will have missing values for some variables. One way to deal with the missing values is to create complete data via various data imputation procedures so that the data can be analyzed as a complete dataset. One of the problems resulting from the imputation of the missing data is that a single imputation (SI) may lead to underestimation of the variances because the variance due to the data imputation has not been accounted for (Rubin, 1987). Multiple imputation (MI) has been favored by many researchers over a single imputation in recent years because it allows people to estimate the imputation variance and, as a result, to use variance estimates adjusted for imputation in the analysis of data involving imputed values (Ghosh-Dastidar and Schafer, 2003; Cole, et al., 2006; Reiter and Raghunathan, 2007; Drechsler, et al., 2008). Although MI is theoretically better than SI, it demands more resources to produce, manage, and analyze the data. There may not be much reason to do MI if the variance underestimation by SI is minimal.

To date, the NHDS has used only a single imputation for its imputed values with no attempt to measure the variation due to that imputation. The purpose of this research is to

[1] The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Center for Health Statistics or the Centers for Disease Control and Prevention.

explore MI as a tool to estimate the imputation variances for the point estimates of the National Hospital Discharge Survey (NHDS). Variance estimates were produced for the imputation procedure applied specifically to the 2006 NHDS data in an attempt to answer the following two questions:

1. Has the SI used in 2006 NHDS led to an underestimation of the variance of the survey estimates?
2. If the answer to the first question is yes, then how much is this underestimation of the variance of the survey estimates?

Information from this research may help data users to better analyze the NHDS data and interpret the analytic results, help the NCHS staff decide whether a MI procedure should replace the current SI procedure, and add to general understanding of SI and MI in a complicated national survey.

1.2 An introduction of NHDS and its hot-deck data imputation

NHDS is conducted by the National Center for Health Statistics to produce nationally representative estimates of the characteristics of discharges, lengths of stay, diagnoses, surgical and non-surgical procedures, and patterns of use of care in U.S. hospitals (DeFrances, Cullen, and Kozak, 2007). The survey uses a complex sample of discharges from hospitals. Data about the sampled discharges are abstracted from inpatient hospital records. A detailed description of the design and development of the NHDS is included in Dennison and Pokras (2000). Only three variables, namely age, sex, and length of stay (LOS), are imputed using a hot deck imputation procedure.

The hot deck procedure assumes that data are missing at random and replaces missing values with the values found in records randomly selected from a pool of similar, but complete, records in the same data set (Rubin, 1987; Kim and Fuller, 2004). See Section 2.1 and Table 1 for how the pool of similar records for the missing values is defined in the NHDS. When used properly, this data imputation procedure retains the distribution of the variable among respondents included in the pool of similar records, allowing the resulting dataset to be analyzed by complete-data methods. This procedure selects the imputed values based on a random procedure. A particular set of imputed values is only one possible sample of values. The use of single imputation ignores the sampling variability and, thus, tends to underestimate the variance (Rao and Shao, 1992). This problem can be resolved by using MI procedures (Rubin, 1987; Rubin and Schenker, 1991).

Currently the hot deck imputation procedure used in the NHDS is an SI procedure. It is not known whether the current variance approximation has resulted in major underestimation of the variance. Using MI as a tool, underestimation of the variance was detected for fewer than half the NHDS statistics included in this study and the magnitude of that underestimation appeared relatively minimal for most of those statistics.

2. Methodology

2.1 Hot deck imputation in 2006 NHDS

Data from the 2006 NHDS were used in this analysis. The missing values for three variables were imputed independently using a hot deck procedure. These three variables were age, gender, and LOS, the only variables imputed in the NHDS. There were a total of 378,579 records in the 2006 NHDS. The number of missing values were 1272 (0.34%), 453 (0.12%), and 829 (0.22%) for age, gender, and LOS, respectively (Table 1). For the variables age and gender, each missing value was imputed with the value found

for that variable in a record randomly selected from all complete-data records for which the principal diagnosis (DX1) matched the DX1 reported in the record with the missing value. The imputation was completed in a single matching step because there were no missing values for DX1. Similarly, LOS was imputed from records in which the principal procedures (PD1) matched the PD1 reported in the record with the missing LOS. Because 37% of the records did not have any procedures (which happens if the discharge associated with the record is for a medical admission without any procedures), only 73% of the missing LOS values could be imputed on the first attempt. The remaining LOS missing values were imputed from records that matched on DX1 (Table 1).

Table 1. Variables imputed, the corresponding variables used in matching to identify similar records for hot-deck imputation, and percent of records missing values for those variables: 2006 NHDS

Variables imputed			**Variables used in matching to identify similar records for hot-deck**		
Name	**Description**	**Missing rate**	**Name**	**Description**	**Missing rate**
AGE	Patient age in years	0.34%	DX1	Principal diagnosis code	0%
SEX	Patient sex	0.12%	DX1	Principal diagnosis code	0%
LOS	Length of hospital stay in days	0.22%	PD1	No Principal procedure	36.55%
			DX1	Principal diagnosis code	0%

2.2 The MI procedure and variance computation

The imputation done for data production was regarded as the first imputation. Six additional imputation cycles were conducted so that the total number (m) of imputations was seven. Equations (1), (2), (3), and (4) below outline the major analytical steps based on the multiple imputations performed in this study, which essentially follows the methodology described by Rubin (1987) and Rubin and Schenker (1991):

$$\bar{Q} = \frac{1}{m}\sum_{t=1}^{m} Q_t \tag{1}$$

$$B = \frac{1}{m-1}\sum_{t=1}^{m}\left(Q_t - \bar{Q}\right)^2 \tag{2}$$

$$\bar{U} = \frac{1}{m}\sum_{t=1}^{m} U_t \tag{3}$$

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B \tag{4}$$

$$\nu = (m-1)\left[1 + \frac{\bar{U}}{\left(1+m^{-1}\right)B}\right]^2 \tag{5}$$

In equations (1), (2), (3), (4), and (5),

m is the number of imputation cycles performed,

Q_t is the complete-data point estimate of the 2006 NHDS data *from imputation cycle t*,

U_t is the complete-data variance estimate from *imputation cycle t*,

$\bar{Q}$ is the mean of the complete-data point estimates from multiple imputations,

B is the variance between the imputations,

$\bar{U}$ is the average variance within the imputations,

T is the total variance, i.e. the variance for $\bar{Q}$, and

ν is the degrees of freedom for significance tests and interval estimates of $\bar{Q}$.

The estimates used in this study were selected from the estimates in Table 1 of a report on the 2006 NHDS by DeFrances, et al. (2008). They are the aggregate estimates for the number of patient discharges and the average LOS by sex, age groups and region. Comparisons are made between the SI and the MI for the estimates and their corresponding standard errors (SE). The change in SE reflects the change in the variance because SE is the square root of the variance for the error term by definition.

2.3 Measurement of MI effects as compared to SI

If the imputation variance is large, then the variance obtained from the MI procedure in equation (4) should be significantly greater than that from the SI procedure (the control). To compare the point estimates between the MI and SI, the percentage differences between the MI and SI estimate were calculated using the following formula:

$$\text{Change\%} = 100 \times \frac{SI - MI}{SI} \tag{6}$$

where SI is the estimate or the SE from the SI for the statistic of interest, and MI is the estimate or the SE for the same statistic from the MI. A negative value of Change% indicates that the SI estimate or SE is smaller than the MI estimate or SE and therefore the SI may have underestimated the estimate or the SE. The greater the absolute value of the Change%, the greater is the effect of MI.

The magnitude of the difference

$$\text{Change} = \text{SI-MI} \tag{7}$$

was also calculated and examined for estimates and SEs. Significance tests of the difference $\left(\hat{Q}_1 - \bar{Q}\right)$ were not performed because it is currently not known how to correctly compute the covariance needed in the variance of $\left(\hat{Q}_1 - \bar{Q}\right)$ for the t-test statistic.

3. Results and discussion

3.1 Imputation estimates and variances

The discharge and LOS estimates from SI and MI and the corresponding SE by sex, region and age groups are presented in Table 2. Table 2 is divided into three sub-tables, i.e. Tables 2A, 2B, and 2C, which contain the data for males and females combined, males only, and females only, respectively. Because estimates of total numbers of discharges for the nation and regions are not affected by the imputation, they were excluded from the study and from Table 2A. A total of 50 statistics were included in the study.

Only very small proportions, ranging from 0.12% to 0.34%, of the data were missing for the three imputed variables in the 2006 NHDS (Table 1). Therefore large differences between the MI estimates and the SI estimates were not expected. The Change% results

were small in most cases (see Table 2). The Change% for estimates ranged from -1.046 to 2.196 while the Change% for SEs ranged from -2.449 to 4.058. Of the 50 statistics included in the study, 45 (90%) estimates and 41 (82%) SEs had absolute values for Change% that were less than one percent. On the other hand, the absolute change for aggregate discharge estimates was greater than one thousand for 14 of the 23 study aggregate estimates and ranged up to almost 49 thousand while the maximum absolute change in SEs for the aggregate statistics was about 11 thousand. (Because aggregate estimates are rounded to thousands in publications, changes of one thousand discharges would affect the published estimates.) For the 27 LOS statistics in the study, the maximum absolute values of changes between SI and MI was only 0.055 days for point estimates and only 0.008 days for SEs (not shown). Thus, it appears the effect of MI on LOS is negligible.

If the SI underestimated the variance, then the Change% values would have a negative sign. This was not the case for most of the study statistics. Of the 50 Change% values for the SE, 22 (or 44%) had negative signs (Table 2). Of those 22 SEs, only three had absolute Change% values greater than 1% and only four had absolute changes exceeding one thousand discharges. Those four with changes exceeding one thousand were for total discharges to males, discharges to males in the South, discharges to males 15-44 years old, and total discharges to patients 15-44 years old.

The observation that so many (44%) of the SI estimates of SE exceed the MI estimates of SE was intuitively unexpected, but it is plausible. From Section 2.2, the total (MI) variance is the sum of two parts: the average of variances within imputation and the variance between imputations. Probably because the portion of data imputed for each variable is small in the NHDS, the MI variances are dominated by the within imputation variances (see examples in Table 3). Hence, when the SI estimate of SE exceeds that from MI, the within variance for the SI exceeds the within variances for some, if not all, the other imputation cycles used in the MI. Because both SI and MI yield different results every time they are used, the event in which an SI estimate for SE exceeds that from MI is random.

It appeared that the MI had the greatest effect on statistics involving patients who are under 15 years of age. Among each group of study statistics defined by gender (both, male, female) and statistic type (discharge count or LOS), the maximum absolute Change% values and absolute changes occurred for statistics involving patients under 15 year of age except for the SEs of statistics about LOS for males. Also, the maximum Change% for SEs was 4.058, a positive value, which occurred for estimated discharges to males under 15 years of age, and the maximum Change in SE was 11 thousand which occurred for total discharges to patients under 15 years of age (Table 2A).

The MI appeared to have more effect on study statistics which involve age groups than estimates which involve regions. For Change% of estimates, the values ranged from -1.046 to 2.196 for estimates by age group while the corresponding range was -0.094 to 0.101 for estimates by region. For Change% of SEs, the values ranged from -2.449 to 4.058 for estimates by age group while that range was -0.842 to 1.283 estimates by region. For magnitudes of change in estimates, the maximum was 49 thousand for estimates by age and 5 thousand for estimates by region. The magnitudes of change in SEs ranged up to 11 thousand for estimates by age and up to 1 thousand for estimates by region.

The MI appeared to have minimal effect on LOS statistics. While the range of Change% values was -0.462 to 1.132 for LOS estimates and -0.842 to 1.895 for SEs of those estimates, the maximum absolute change was only 0.055 days for LOS estimates and 0.008 days for their SEs (not shown).

Table 2. Estimates produced with single imputation (SI) and multiple imputation (MI) for discharges and the average length of stay (LOS) by sex, age groups and region: 2006 NHDS. (All estimates shown are produced using survey weights. Estimates of total numbers of discharges for the nation and regions are excluded. Values shown for Change% may differ from those derived from the shown SI and MI values because of rounding.)

Table 2A. Males and females combined

Items	Estimated values			Standard error		
	SI	MI	Change%	SI	MI	Change%
Number of discharges (in thousands)						
Under 15 years	2,298	2,249	2.129	368	357	2.937
15-44 years	10,800	10,848	-0.446	437	441	-0.868
45-64 years	8,686	8,690	-0.042	320	321	-0.057
65 years and	13,070	13,068	0.022	507	506	0.005
Average length of stay in days (LOS)						
Total	4.773	4.772	0.007	0.0669	0.0668	0.108
Under 15 years	4.783	4.758	0.529	0.2480	0.2512	-1.308
15-44 years	3.749	3.759	-0.279	0.0838	0.0835	0.414
45-64 years	4.989	4.988	0.019	0.0856	0.0857	-0.067
65 years and	5.473	5.472	0.013	0.0710	0.0707	0.433
Northeast	5.285	5.286	-0.027	0.1230	0.1229	0.121
Midwest	4.224	4.225	-0.024	0.0973	0.0973	-0.008
South	4.907	4.906	0.026	0.1017	0.1016	0.099
West	4.599	4.596	0.044	0.2096	0.2093	0.156

Table 2B. Males only

Items	Estimated values			Standard error		
	SI	MI	Change%	SI	MI	Change%
Number of discharges (in thousands)						
Total	13,990	13,994	-0.025	548	549	-0.208
Under 15 years	1,295	1,266	2.196	209	200	4.058
15-44 years	2,922	2,952	-1.046	148	151	-1.769
45-64 years	4,287	4,291	-0.100	167	168	-0.351
65 years and	5,487	5,484	0.053	216	216	0.065
Northeast	3,045	3,044	0.024	233	233	0.025
Midwest	3,136	3,137	-0.008	331	331	-0.007
South	5,220	5,225	-0.094	275	277	-0.683
West	2,589	2,588	0.034	244	245	-0.184

Table 2B. Males only (continued)						
Items	Estimated values			Standard error		
	SI	MI	Change%	SI	MI	Change%
Average length of stay in days (LOS)						
Total	5.181	5.180	0.017	0.078	0.078	0.714
Under 15 years	4.875	4.820	1.132	0.219	0.221	-0.874
15-44 years	4.988	5.011	-0.462	0.133	0.131	1.630
45-64 years	5.107	5.107	0.000	0.098	0.098	-0.597
65 years and	5.414	5.412	0.042	0.079	0.077	1.895
Northeast	5.646	5.647	-0.023	0.141	0.141	0.254
Midwest	4.394	4.396	-0.043	0.099	0.099	0.128
South	5.407	5.406	0.033	0.104	0.103	0.450
West	5.132	5.126	0.101	0.271	0.267	1.283

Table 2C. Females only						
Items	Estimated values			Standard error		
	SI	MI	Change%	SI	MI	Change%
Number of discharges (in thousands)						
Total	20,864	20,860	0.017	765	764	0.117
Under 15 years	1,003	983	2.041	160	158	1.471
15-44 years	7,878	7,896	-0.223	326	327	-0.229
45-64 years	4,399	4,399	0.014	164	163	0.397
65 years and	7,584	7,584	0.000	300	300	-0.023
Northeast	4,232	4,232	-0.017	281	281	-0.017
Midwest	4,815	4,814	0.005	520	520	0.004
South	7,920	7,915	0.062	400	399	0.340
West	3,898	3,899	-0.022	265	265	0.208
Average length of stay in days (LOS)						
Total	4.499	4.499	0.002	0.064	0.064	-0.344
Under 15 years	4.665	4.678	-0.288	0.343	0.351	-2.449
15-44 years	3.289	3.291	-0.060	0.062	0.062	0.240
45-64 years	4.874	4.872	0.040	0.096	0.096	0.363
65 years and	5.515	5.516	-0.007	0.080	0.081	-0.886
Northeast	5.024	5.026	-0.031	0.117	0.117	0.002
Midwest	4.114	4.114	-0.010	0.116	0.116	-0.051
South	4.577	4.576	0.031	0.106	0.106	-0.174
West	4.244	4.245	-0.005	0.172	0.174	-0.842

3.2 Degrees of freedom for MI variance

The standard errors of MI estimates and the degrees of freedom for the MI variance estimates were calculated using equations (4) and (5). Components of SEs for MI estimates of discharges by age groups are presented in Table 3 to demonstrate the relative magnitude of the respective values. The degrees of freedom calculated for variances were

all very large owing to the fact that the within standard error, $\sqrt{\overline{U}}$, is much greater than the between standard error, $\sqrt{B}$ (Table 3).

Table 3. Standard errors and degrees of freedom for SE of MI estimates of total discharges by age group.

Age group	SE computation for MI (in thousands)			Degrees of freedom for SE of MI estimate
	Within MI	Between MI	Total = SE for MI estimate	
Under 15 years	356.8	1.8	358.9	7.032E+09
15-44 years	440.6	0.6	441.3	1.078E+12
45-64 years	320.6	0.1	320.7	3.195E+14
65+	506.5	0.1	506.6	6.072E+15

4. Conclusions

The potential effect of multiple imputation (MI) on estimates and variances from the 2006 National Hospital Discharge (NHDS) was explored. The missing values for three variables (age, sex, and LOS) are imputed in the NHDS and to date, only a single imputation (SI) has been used for that imputation. For this study, single and multiple imputation estimates and their standard errors were compared in terms of change and percent change from the values which were based on single imputation for 50 statistics typically published in reports based on NHDS data. These statistics were for discharges and average LOS for all patients, for four age groups (<15 years, 15 to 44 years, 45 to 64 years, and 65 years or more), and for four regions (Northeast, Midwest, South, and West).

The percentage of records with missing values was very small (<0.4%) for each of the three imputed variables. As a result, MI was not expected to have a large impact on the survey estimates and their corresponding standard errors. In general MI did have little effect as expected. For the study statistics, the percent changes caused by MI were less than one percent for 90% of the estimates and 82% of the SEs. The effect of MI on LOS study statistics was minimal. However, the effects of MI caused changes in excess of 1% of the corresponding SI value and/or caused notable changes in magnitudes of estimates or SEs for some statistics.

The following conclusions can be drawn from the results of this research:

1. Variances due to use of SI were less than those from MI for only 44% of the study statistics. The magnitude of that underestimation was negligible for LOS statistics but was greater than one thousand discharges for four of the aggregate statistics. However, the published point estimates for aggregate statistics would have been affected by use of MI instead of SI for at least 14 (61%) of the 23 aggregate discharge statistics included in the study because the absolute differences between the SI and MI for those point estimates exceeded one thousand.
2. Imputation has the greatest effects on aggregate discharge estimates for the age group “Under 15 years”. Changes of magnitude of 48 thousand are possible in

some estimates for this age group. For SEs, the absolute percent of change from SI caused by MI can be as large as 4.1% and the magnitude of the change can range up to 11 thousand.

3. In general the MI had little effect ($\leq 1.3\%$) on both regional estimates and their corresponding SEs. In certain cases, however, it can change some published discharge estimates by a magnitude of 5 thousand.

One needs to keep in mind the above conclusions are based on the 2006 NHDS only and on the method currently used by NDHS to impute missing values. This research does not include alternate procedures which could be used for imputing those values.

References

Cole, S.R., Chu, H., and Greenland, S. (2006), "Multiple Imputation for Measurement-Error Correction," *International Journal of Epidemiology*, 35, 1074–1081.

DeFrances, C.J., Lucas, C.A., Buie, V.C., and Golosinskiy, A. (2008), "2006 National Hospital Discharge Survey", *National Health Statistics Reports*, Number 5, July 30, 2008.

Dennison, C.F., and Pokras, R. (2000), "Design and operation of the National Hospital Discharge Survey: 1988 redesign", National Center for Health Statistics. Vital and Health Stat 1(39).

Drechsler, J., Dundler, A., Bender, S., R¨assler, S., and Zwick, T. (2008), "A New Approach for Disclosure Control in the IAB Establishment Panel – Multiple Imputation for a Better Data Access," *Advances in Statistical Analysis*, 92, 439–458.

Ghosh-Dastidar, B., and Schafer, J.L. (2003), "Multiple Edit/Multiple Imputation for Multivariate Continuous Data", *Journal of the American Statistical Association*, 98, 807–817.

Kim, J.K., and Fuller, W. (2004), "Fractional Hot Deck Imputation", *Biometrika*, 91, 559-578.

Rao, J.N.K., and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation," *Biometrika*, 79, 811–822.

Reiter, J.P., and Raghunathan, T.E. (2007), "The Multiple Adaptations of Multiple Imputation," *Journal of the American Statistical Association*, 102, 1462-1471.

Rubin, D.B. (ed.) (1987), "Multiple Imputation for Nonresponse in Surveys", New York: JohnWiley & Sons, pp 1-23.

Rubin, D.B., and Schenker, N. (1991), "Multiple Imputation in Health-Care Databases: An Overview and Some Applications," *Statistics in Medicine*, 10, 585–598.