Variances of Imputation Variances as Determiner of Sufficient Number of Imputations Using data from 2012 NAMCS Physician Workflow Mail Survey [1]

Qiyuan Pan, Rong Wei, Iris Shimizu, Eric Jamoom
National Center for Health Statistics, Centers for Disease Control and Prevention, USA
Email: qpan@cdc.gov

ABSTRACT

This paper presents an empirically-based method for a data user to determine the sufficient number of imputations for his or her data. The minimally sufficient number of imputations was determined based on the relationship between *m*, the number of imputations, and ω, the standard error of imputation variance estimates using the 2012 National Ambulatory Medical Care Survey (NAMCS) Physician Workflow mail survey.

Keywords: Multiple imputation, sufficient number of imputations, hot-deck imputation

1. Introduction

Multiple imputation (MI) has been steadily gaining popularity in the past several decades. Researchers have different recommendations on how many imputations are sufficient in applying MI. Rubin suggested 2 to 5 [1] [2]. Schafer and Olsen suggested 3 to 5 [3]. Graham et al. suggested 20 or more [4]. Hershberger and Fisher suggested that several hundred imputations are often required [5]. Allison suggested that one may need more imputations than what were generally recommended in the literature [6]. It may not be practical to recommend a specific number of imputations that can universally fit all imputation models or data and management situations. Instead of attempting to prove or disprove recommendations found in the published literature, the current research examines an empirically-based method to determine the minimally sufficient number of imputations needed.

Let *m* be the number of imputations and η be the least number of imputations that is sufficient to meet a minimum requirement chosen by a data analyst. Rubin [1] established that the relationship between $V(Q_m)$, the large sample variance of a point estimator, Q, from a finite *m,* and $V(Q_\infty)$, the variance from an infinite *m,* is

$$V(Q_m) = \left(1 + \frac{\gamma_0}{m}\right) V(Q_\infty), \tag{1}$$

where γ_0 is the population fraction of missing information. From this relationship, the relative efficiency (RE) measured in the units of standard errors from using MI is

$$RE = \left(1 + \frac{\gamma_0}{m}\right)^{-1/2}. \tag{2}$$

Based on this RE, Rubin stated: "If $\gamma_0 \leq 0.2$, even two repeated imputations appear to result in accurate levels, and three repeated imputations result in accurate levels even when $\gamma_0 = 0.5$".

---

[1] The findings and conclusions in this paper 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 of the United States government.

The number of imputations affects the MI results in multiple ways as indicated by the following equations for MI data analyses [1]. The mean of the point estimator Q is

$$\overline{Q} = \frac{1}{m}\sum_1^m Q_i. \quad (3)$$

The average variance estimate of Q over *m* complete datasets from MI is

$$\overline{U} = \frac{1}{m}\sum_1^m U_i, \quad (4)$$

where U_i is the variance estimate of the ith imputation. The estimated imputation variance is

$$B = \frac{1}{m-1}\sum_1^m (Q_i - \overline{Q})^2. \quad (5)$$

The total variance estimate is

$$T = \overline{U} + (1 + m^{-1})B. \quad (6)$$

Equations (3) to (6) all contain *m* as a factor, indicating that *m* can affect the results of MI data analyses in multiple ways. The minimally sufficient number of imputations could be defined as the smallest *m* that would produce a sufficiently accurate $\overline{Q}$, $\overline{U}$, B, or T as judged by the data user.

This paper presents a methodology to determine η by focusing on the effects of *m* on the accuracy of B as defined by equation (5). The focus is on B because the primary advantage of MI over single imputation (SI) is that MI makes it possible to estimate B while SI cannot [1] [7]. For an *m* to become the η, this *m* should first be capable of allowing us to obtain a sufficiently accurate B. One way to judge the accuracy of B is by examining ω, the estimated standard error of B. If ω becomes sufficiently small, then one may conclude that B is sufficiently accurate and the *m* corresponding to that ω value would represent η.

2. Methodology

2.1. The survey

Data from the 2012 wave of the National Ambulatory Medical Care Survey (NAMCS) Physician Workflow study (PWS) were used to illustrate the proposed methodology for determining the minimally sufficient number of imputations needed. The PWS is a nationally representative, 3-year panel mail survey of office-based physicians conducted by the National Center for Health Statistics (NCHS) [8].

The PWS sample includes 5,266 physicians who were confirmed eligible in the 2011 Electronic Medical Records mail survey, a supplement to the NAMCS. To meet eligibility criteria, physicians had to see ambulatory patients in office-based settings. All eligible physicians were mailed the first wave of the study in 2011. Sampled physicians who did not respond to the first mailing were sent up to two additional mailings, and survey administrators followed up by phone with those who did not respond to any of the three mailings. A total of 3,180 eligible physicians responded in the 2011 PWS, yielding a weighted response rate of 46.0 percent. All eligible physicians, including those who did not respond in 2011, were mailed a second survey in 2012. This 2012 cycle of data yielded 2,567 eligible responses, for a weighted response rate of 42.1 percent and was used for the current MI research. Missing value percentages were calculated by regarding the 2,567 records as the complete data set and the imputations were carried out within these 2,567 records. All analyses in this research were conducted with unweighted data.

2.2. The imputed variables

The survey variables to be imputed in this study are described in Table 1. Results of two imputed variables, PRACSIZE5 and CLSTAFF1, are included in this paper. PRACSIZE5 was the number of physicians in the practice. The valid values for the number of physicians in the practice ranged from 1 to 100 in the raw survey data but they were recoded into five categories for PRACSIZE5. CLSTAFF1 was the number of clinical staff with 8.88% missing data and 0-99 value range. These two variables were selected because they differed in variance and the percentage of missing data.

2.3. The imputations

Hot deck imputation [9] was used in this MI study. The donor groups for the hot deck imputation were defined by the region of the physician's interview office (REGION), the physician's interview specialty group (SPECR), and the physician's primary present employment code (PRIMEMP) (Table 1). These three variables were chosen for defining the donor groups in order to minimize possible correlation between the donor groups and the variables to be imputed. The donors for missing values were randomly selected with replacement from the donor group which matched the recipient. A total of 500 imputations were obtained for each variable. The pool of 500 imputations was used as the population and a random sample of *m* imputations was drawn from the pool for the MI.

Table 1. Characteristics of variables used

Variable name	**Description**	**Possible values**	**Missing (%)**
The imputed variables			
PRACSIZE5	Practice size: The number of physicians at the reporting location grouped into 5 categories	1 to 5	3.62
CLSTAFF1	Number of clinical staff	0 to 99	8.88
The independent variable used in hot-deck imputation			
REGION	Region of the physician's interview office	1 to 4	0
SPECR	Physician specialty	1 to 15	0
PRIMEMP	Primary present employment code	11,13,20,21, 22,23,30,31	0

2.4. Determination of the variance of the imputation variances

Thirteen different numbers of imputations were tested for MI: m=2, 3, 5, 10, 15, 20, 25, 30, 35, 40, 60, 80, and 100. The imputation variance (B), defined by equation (5), was calculated for the data obtained from each MI. In order to calculate the variance of B, 10 independent random samples from the 500 imputations were pulled for each m of each variable. Figure 1 describes the process in a diagram. V_B, the variance of B, is estimated by equation (7):

$$V_B = \frac{1}{(n-1)} \sum_1^n (B - \overline{B})^2, \qquad (7)$$

where n is the number of MI samples for a given m, which is 10 for the current study. The standard deviation of B is:

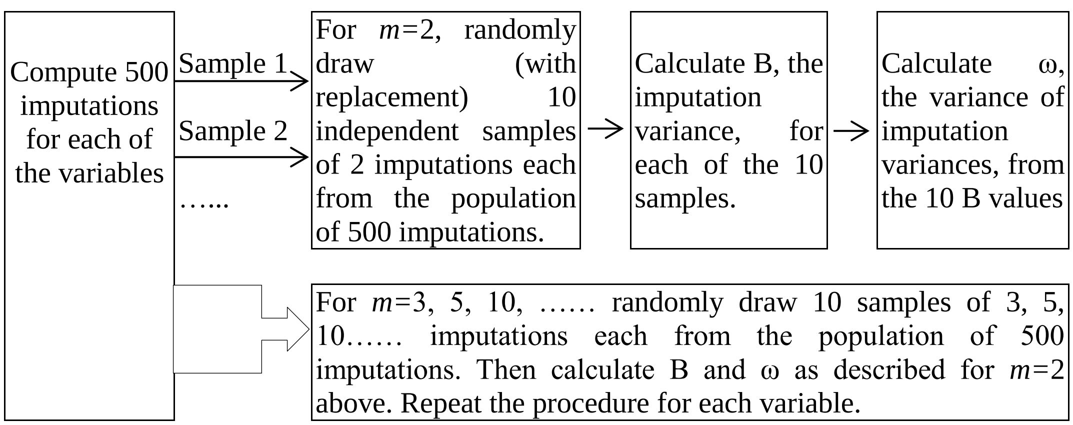
$$SD_B = \sqrt{V_B} = \sqrt{\frac{1}{(n-1)} \sum_1^n (B - \overline{B})^2} \qquad (8)$$

The standard error of B was calculated using the following formula:

$$\omega = \sqrt{\frac{V_B}{n}} = \sqrt{\frac{1}{n(n-1)} \sum_1^n (B - \overline{B})^2} \qquad (9)$$

V_B, SD_B, and ω measure the variance of imputation variances at different scales.

Figure 1. Experimental design for obtaining B, the imputation variance, and ω, the standard error of B, for different m, the number of imputations.



2.5. Determination of sufficient number of imputations (η)

For a given m-ω curve, different data users may have different criteria for a particular m value to be recognized as being “sufficient” and so arrive at with different η values. As an example, what is called “the confidence interval method” was developed and presented here. In this method, ω is used to calculate the 95% confidence interval for $\overline{B}$, the mean of the 10 B values of each m. Because $\overline{B}$ is the sample mean of B which is a chi-square variable with m-1 degrees of freedom, the normal distribution can be used to approximate the distribution of $\overline{B}$ for sufficiently large values of m [10]. For PRACTSIZE5, m = 2 was verified to be sufficiently large for $\overline{B}$ to approach normality (not shown). As a result, for a sample size of 10, the Student t test can be used to calculate the confidence interval of $\overline{B}$. Then P, the percentage of half-width of the confidence interval divided by $\overline{B}$ is derived and used as the parameter to determine η. P is defined by the following equation:

$$P = 100\ (t_{0.05}\ \omega)/\overline{B}, \qquad (11)$$

where $t_{0.05}$ is the t value from the Student t test at 0.05 probability level and the term "$t_{0.05}$ ω" represents half the width of the 95% confidence interval because the full width of the 95% confidence interval would be $[\overline{B} + (t_{0.05}\ \omega)] - [\overline{B} - (t_{0.05}\ \omega)] = 2\ t_{0.05}\ \omega$. The variable PRACSIZE5 was used to illustrate this method and the results are presented in Table 2. Data in Table 2 show that P decreased as *m* increased. It is up to each data analyst to decide the cut-off point at which the *m* would be recognized as sufficiently large so that this *m* would become η. The minimum *m* value above the chosen cutoff point of P would be the η. For illustration, then η is determined to be 40 for PRACTSIZE5 if P=15 is chosen as the cutoff point.

Table 2. The worksheet of the confidence interval method for determination of the sufficient number of imputations (η) using PRACSIZE5 as an example.

m	$\overline{B}$ [a]	**ω**	**Half confidence interval ($t_{0.05}$*ω)**	**Percentage of half confidence interval over $\overline{B}$ (100*$t_{0.05}$*ω / $\overline{B}$)**
2	2.395	0.613	1.386	57.875
3	1.290	0.384	0.868	67.326
5	1.048	0.254	0.575	54.840
10	1.030	0.166	0.375	36.446
15	0.876	0.110	0.249	28.400
20	0.979	0.101	0.229	23.360
25	1.085	0.108	0.244	22.495
30	1.084	0.083	0.189	17.422
35	0.912	0.061	0.138	15.176
40	1.033	0.067	0.151	14.582
60	0.864	0.034	0.077	8.951
80	0.964	0.042	0.095	9.858
100	1.024	0.031	0.070	6.876

[a] $\overline{B}$ = The mean of B (imputation variance).

3. Results and discussions

Individual B values at each *m* were plotted in Figure 2. Within each variable, the individual B values were scattered across a wide range when *m* was small (< 20) (Figure 2). For example, for variable PRACSIZE5 at *m*=3, the B value was 0.2186 from one sample, and 4.0896 from another sample. This means that if one decided to use *m*=3 for MI, the imputation variance could by chance be many times bigger or smaller than the true B value. The widely scattered data points at low *m* values in Figure 2 indicate that when *m* was less than 20 or even 40, the B value obtained may not be accurate and reliable.

Figure 3 presents data on ω for different *m* values. Measured in standard error, ω was the largest when *m*=2 or 3. It quickly decreased with the increase in *m* and tended to stabilize as *m* approached 100, the highest *m* tested. Taking the variable PRACSIZE5 as an example, ω decreased by 82% when *m* increased from 2 to 15, then decreased an additional 13% when *m* increased from 15 to 100. Sufficient numbers of imputations (η) as determined by the confidence interval method (see Section 2.5) were 60 for PRACTSIZE5 and 80 for CLSTAFF1.

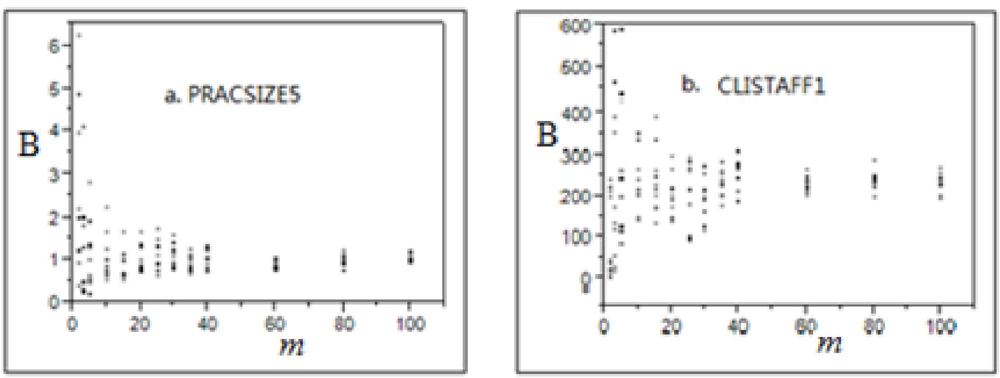


Figure 2. Effects of the number of imputations (*m*) on imputation variances (B)

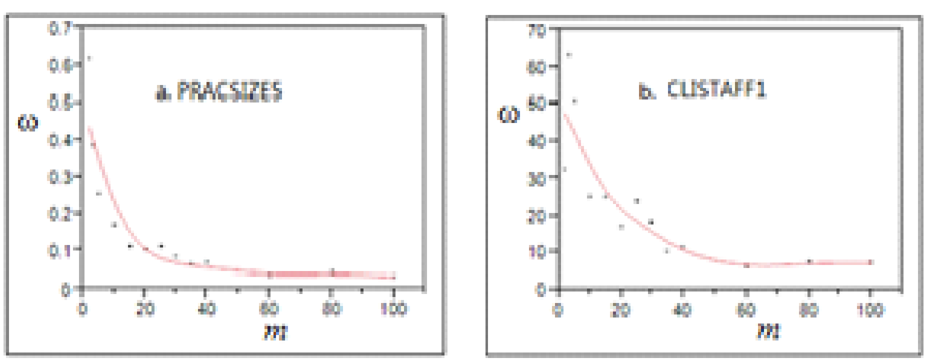


Figure 3. Effects of the number of imputations (*m*) on the variance of the imputation variances (ω)

Shifting from SI to MI allows an estimate of B, the imputation variance, and a more accurate estimate of the total variance, T, as defined by equation (6) [1] [7]. If an MI protocol cannot give a reliable estimate of B, then the major benefit of MI would be lost. Therefore, it makes sense to use an *m*-dependent measure of reliability for the estimate of B as a criterion in determining η, the minimally sufficient number of imputations. The results of this research indicate that B values obtained for $m < 20$ may not be reliable (Figures 2 and 3). If one uses Rubin's recommendation of $m = 2$ or 3, the B value obtained could be many times bigger or smaller than the true B value.

The variance of imputation variances is a good determiner of η for three reasons. First, it is critical for any MI procedure to produce a reliable estimate of B. Second, the effect of m on ω can be big and can be easily visualized, as shown in Figure 3. Furthermore, the effect of m on ω decreases with increased m, allowing the data user to set a cutoff point for an m to be recognized as the η. Third, the method of calculating ω is relatively simple and practical.

To determine the minimally sufficient number of imputations, one has to first define the criteria for being recognized as "sufficient". Different data users may have different criteria. As a result, different η values may be obtained from the same m-ω relationship data. This is another reason why a universal recommendation of η is not a good idea and may even be impossible.

4. Conclusions

The reliability of B can be measured by ω, the variance of imputation variances. Data in this study indicate that ω decreased as m increased and the unit gain from increased m decreased with greater m. The m-ω curve can be used to determine η, the minimally sufficient number of imputations in MI. The method described in this paper can be used by any data user to determine the η that fits his or her particular data situation. The most popular recommendation for η is between 2 and 5, suggested by Rubin [1] [2]. Our results indicate that 2-5, or even 10, imputations may not be sufficient to obtain a statistically reliable B in MI data analyses.

5. Acknowledgement

The NAMCS Physician Workflow mail survey is sponsored by the Office of the National Coordinator for Health Information Technology.

REFERENCES

[1] D.B. Rubin, "Multiple Imputation for Nonresponse in Surveys," New York: John Wiley & Sons, 1987, pp. 1-23 and pp. 75-147.

[2] D.B. Rubin, "Multiple imputation after 18+ years (with discussion)," Journal of the American Statistical Association, Vol. 91, No. 434, 1996, pp. 473-489.

[3] J. L. Schafer and *m*. K. Olsen, "Multiple imputation for multivariate missing data problems: A data analyst's perspective," Multivariate Behavioral Research, Vol. 33, No. 4, 1998, pp. 545–571.

[4] J. W. Graham, A. E. Olchowski and T. D. Gilreath, "How many imputations are really needed? Some practical clarifications of multiple imputation theory," Prevention Science, Vol. 8, No. 3, 2007, pp. 206–213.

[5] S. L. Hershberger and D. G. Fisher, “A note on determining the number of imputations for missing data. Structural Equation Modeling,” Structural Equation Modeling, Vol. 10, No. 4, 2003, pp. 648–650.

[6] P. Allison, “Why You Probably Need More Imputations Than You Think,” 2012. http://www.statisticalhorizons.com/more-imputations

[7] D. B. Rubin, “Multiple imputations in sample surveys — a phenomenological Bayesian approach to nonresponse,” Proceedings of the Survey Research Methods Section of the American Statistical Association, 1978, pp. 20-34.

[8] E. Jamoom, P. Beatty, A. Bercovitz, et al. “Physician adoption of electronic health record systems: United States, 2011”, NCHS data brief, no 98, Hyattsville, Maryland, USA, National Center for Health Statistics, 2012.

[9] R. R. Andridge1 and R. J. A. Little, “A Review of Hot Deck Imputation for Survey Non-response”, International Statistical Review, Vol. 78, No. 1, 2010, pp. 40–64.

[10] D. R. Freedman, R. Pisani and R. Purves, “Statistics, Fourth Edition,” New York, W. W. Norton & Company, 2007, pp. 415-424, 488-495, and 523-540.