Effects of number of imputations on fraction of missing information in multiple imputation[1]

Qiyuan Pan
National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD20782, USA
Email: qpan@cdc.gov

Abstract

The fraction of missing information (γ) and the number of imputations (m) are two important parameters in multiple imputation (MI). They are used them to define the relative efficiency (RE) of MI: $RE = (1+\gamma/m)^{-1/2}$. Based on this RE, a very influential conclusion was made that a small m (≤5) would be sufficient for MI. However, evidences for much greater m have been accumulating. A better understanding of m-γ relationship is of importance in MI research and application. The effects of m on γ were examined using the data of the 2012 Physician Work Flow Mail Survey, which was a supplement to the National Ambulatory Medical Care Survey data. The results suggest that γ reduces with the increase of m, shaking the foundation of using the γ-based RE to determine the sufficient m.

Key Words: Multiple imputation; Fraction of missing information; Number of imputations; Missing data; National Ambulatory Medical Care Survey

1. Introduction

Multiple imputation (MI) has become a popular approach of handling missing data [1, 2, 3]. Prior to adopting MI, one has to decide the number of imputations (m) that is appropriate for one's particular data situation and analytical needs. What m should be considered sufficient? The most influential answer, which was given by Rubin in his 1987 classic book on MI, is that only a few imputations, i.e. m≤5, would be sufficient [4]. This conclusion was based on the relative efficiency (RE) as defined below:

$$RE = (1 + \frac{\gamma_0}{m})^{-\frac{1}{2}} \quad (1)$$

where γ_0 is the population value of γ, the fraction of missing information. The γ at a finite m, γ_m, is defined as:

$$\gamma_m = \frac{r+2/(v+3)}{r+1}, \quad (2)$$

where r and v are defined as [4]:

$$v = (m-1)(1+\frac{1}{r})^2, \quad (3)$$

$$r = \frac{(1+\frac{1}{m})B}{U}, \quad (4)$$

where B is the between-imputation variance and U is the within-imputation variance, defined by equations (5) and (6) below, respectively [4]:

$$U = \frac{1}{m}\sum_1^m U_i \quad (5)$$

$$B = \frac{1}{m-1}\sum_1^m (Q_i - \overline{Q})^2 \quad (6)$$

---

[1] The statements of this paper do not represent the views of the National Center for Health Statistics (NCHS) or the Centers for Disease Control and Prevention (CDC) of the United States. Dr. Rong Wei of NCHS is acknowledged for her valuable suggestions.

where Q is the quantity of interest, and the subscript *i* denotes the *i*th imputation of the MI [4].

The term "fraction of missing information" sounds similar to fraction of missing data (δ). But actually, γ and δ are very different [5, 6]. δ is a feature of the data and is fixed once the data collection is complete, whereas γ may be affected by whether and how MI is done and how the data are analyzed [5, 6]. The relationship between m and γ is of particular importance because γ was used by Rubin to determine the sufficient m. Nowadays people generally feel that m≤5 is too small and are using much greater m values in their MI [7, 8, 9, 10, 11, 12, 13, 14, 15]. If Rubin's recommendation of small m is not valid, what might have gone wrong? How is γ related to m? How is γ_m related to γ_0? An insight into the m-γ relationship may provide important information for answering these questions. The m-γ relationship has not been explicitly discussed in the published literature. The current research examines the effects of m on γ using the data of 2012 Physician Workflow Mail Survey (PWS12) of the National Ambulatory Medical Care Survey (NAMCS).

2. Multiple imputation trials

The Physician Workflow Mail Survey (PWS) shared the same sampling frame as NAMCS and was considered a supplement to the NAMCS [15]. PWS was conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention, USA. It was a nationally representative, 3-year (2011-2013) panel mail survey of office-based physicians [16], with each year being a complete survey cycle. The data of the 2012 PWS, i.e. PWS12, were used in this research. PWS12 had 2,567 eligible, responded physicians in the sample. The three variables, i.e. SIZE5, SIZE20 and SIZE100, were selected as the variables for imputation (ImpV). They represented the physician's practice size (SIZE). SIZE100 had a value range of 1 to 100. SIZE5 was derived by recoding the values of SIZE100 into 5 categories, and SIZE20 was derived by top-coding the >20 values of SIZE100 into 20. The description of the three imputed variables is in Table 1.

Four levels of δ, i.e. 4%, 10%, 20%, and 29%, were used. PWS12 initially had 29% missing data due to item nonresponse for SIZE. After the missing values were replaced with all the non-missing values of the 2011 data for the same physician, the δ of PWS12 became 4%. The other two δ values, 10% and 20%, were created by partially replacing the missing values in 2012 with the non-missing values in the 2011 survey for the same physician in a random manner. Hot deck imputation [17] was used. The MCAR (missing completely at random) model was assumed and no covariant variables were used in the imputation model. The m values for the MI trials were 3, 5, 10, 20, 30, 40, 60, 80 and 99. Thirty replicates were run for each MI.

The MI data were analyzed by using REGION, PRIMEMP and DERIVED, respectively, as the analytic variable (AnaV), with the control (CONTROL) being the analysis with no analytic variable used. REGION and PRIMEMP are two real variables from PWS12. DERIVED was a derived variable whose values were highly correlated with the ImpV variables. The description of the analytic variables is in Table 2. Analyses were conducted with the un-weighted data. The γ was calculated using equations (2) to (6), with the "quantity of interest", Q, being the means of SIZE5, SIZE20, and SIZE100.

Table 1: The imputation variables (ImpV).

ImpV	Description	Mean	Value range	Total variance
SIZE5	Practice size – 5 group. Derived from SIZE100	3.06	1 – 5	1.97
SIZE20	Practice size – top-coded to 20. Derived from SIZE100	6.47	1 – 20	38.26
SIZE100	Practice size	11.41	1 – 100	483.02

Table 2: The analytic variables (AnaV)

AnaV	Description	Value range
CONTROL	No analytic variable	1 (the whole data)
REGION	Region of the Physicians Interview office	1=Northeast, 2=Mid West, 3=South, 4=West
PRIMEMP	Primary present employment of the physician	14 categories. Value examples: 22=AOA-Office prac group; 40=AMA-Medical school; 64=AMA-County/Cty/State Govt Other
DERIVED	Derived variable whose values are highly correlated with ImpV variables	1 to 4 for SIZE5, 1 to 9 for SIZE20, and 1 to 17 for SIZE 100

3. Results

3.1 γ decreased with increased m

In Figures 1, 2, and 3, the γ values were plotted against the corresponding m values for different combinations of treatment factors δ, ImpV, and AnaV. In Figure 1, the three ImpV variables were included in each graph to examine how ImpV affected the m-γ relationship. In addition, the comparison between graphs a and b and that between c and d allow us to see the difference in m-γ relationship between δ levels, and the comparison between graphs a and c and that between b and d allow us to see the difference between AnaV variables. The general trend was that the γ values decreased as m increased. The magnitude of the γ decrease was larger at smaller m. The ImpV of higher variance such as SIZE100 did not always have greater γ than ImpV of smaller variance such as SIZE5 and SIZE20 (Table 1; Figure 1).

With all four δ levels included in each graph, the primary purpose of Figure 2 was to show the m-γ relationship at different δ levels. A comparison between graphs a and b shows the effects of ImpV and AnaV as well as the interactions among δ, ImpV and AnaV. In general, the decrease of γ with increased m was still obvious when m was less than 30. But exceptions existed. In general, higher δ usually resulted in bigger γ.

Table 3 serves two purposes. First, it numerically shows that the magnitude of the γ decrease with increased m was much greater when m was smaller. Without any exception, the γ value difference between m=3 and m=40 was always much larger than that between m=40 and m=99. Secondly, the fact that γ value at m=40 was almost always greater than that at m=99 suggests that the general trend of γ decrease with increased m continued beyond m>40, which may not be obvious in Figures 1 and 2.

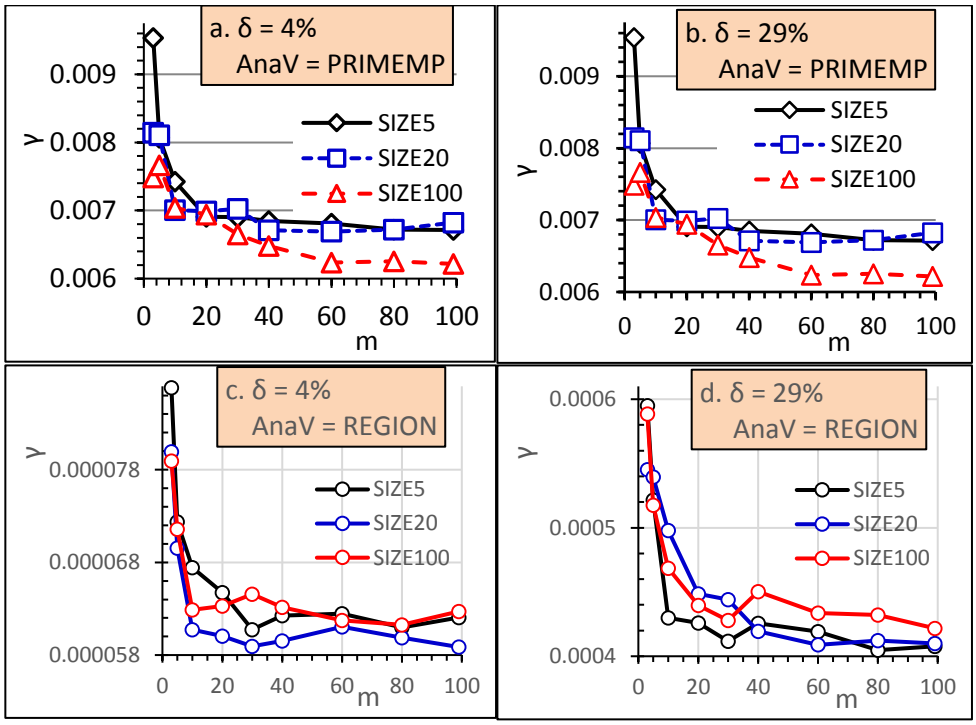

Figure 1. The effects of m on γ for different ImpV variables.

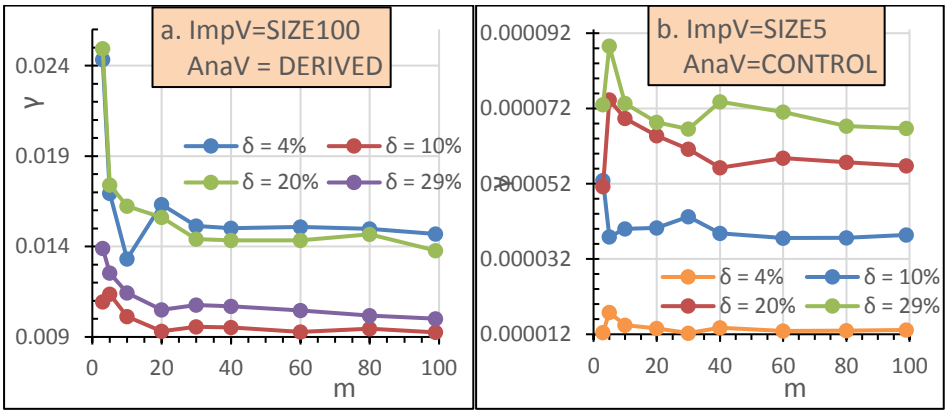

Figure 2. The effects of m on γ for different missing data percentages (δ).

Table 3. Comparison of γ% (=γ×100) between m=3 and m=99 for SIZE5, SIZE20 and SIZE100 for PRIMEMP and δ=4%, 10%, 20%, and 29%.

m	SIZE5				SIZE100			
	δ=4%	δ=10%	δ=20%	δ=29%	δ=4%	δ=10%	δ=20%	δ=29%
3	0.088	0.272	0.960	0.954	0.067	0.246	0.915	0.748
40	0.077	0.214	0.746	0.685	0.057	0.176	0.586	0.648
99	0.074	0.221	0.743	0.671	0.055	0.176	0.627	0.622

3.2 The variation of γ was bigger at smaller m

Figure 3 presents the scatter graph showing the effects of m on γ using the treatment combination of ImpV=SIZE5, AnaV=PRIMIMP and δ=4%. Each dot in this graph represents the γ value of one of the 30 replicates. The dots were more scattered at lower

m, evidencing a greater variation of γ at a smaller m. To numerically examine γ variation at different m values, the coefficient of variation (CV) at different m values was presented in Table 4 for SIZE100 at δ=29 for PRIMEMP. At m=3, the CV was 61.3%. Between m=3 and m=20, CV decreased sharply as m increased. The magnitude of the CV decrease with increased m was smaller as m got bigger (Table 4). The γ value distribution for other treatment combinations was similar.

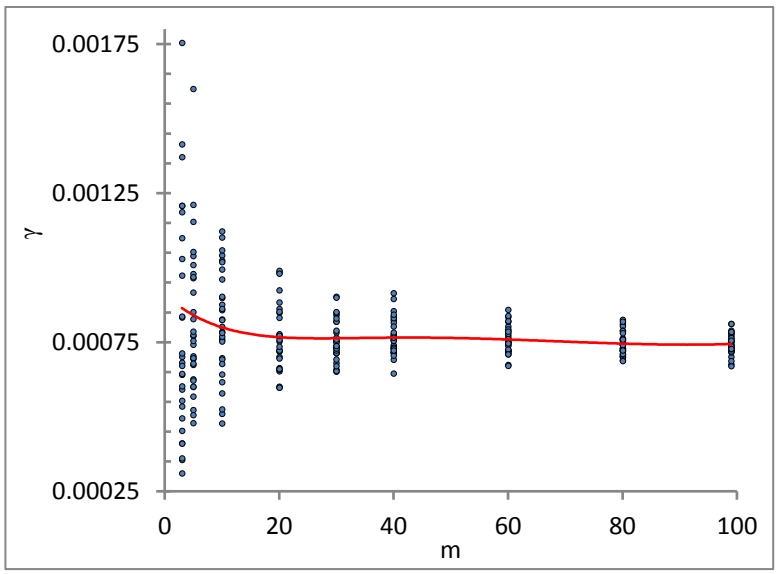

Figure 3. Distribution of the γ values of the 30 replicate samples at different number of imputations (m) for SIZE5 at 4% of missing data percentage for PRIMEMP.

Table 4. The means and the coefficients of variations (CV) of γ at different m values chosen for the MI for SIZE100 at δ=29% for PRIMEMP.

m	Mean γ	CV (%)
3	0.0075	61.3
5	0.0077	45.1
10	0.0070	24.6
20	0.0069	15.8
40	0.0065	10.8
99	0.0062	7.7

4. Discussions

Equations (2) to (6) indicate that γ may be affected by m, B, and U. In the MI trials of this study, we may assume a fixed B and U for the same combination of the treatment factors, i.e. the same δ, ImpV, and AnaV. Figures 1 and 2 suggest that at a fixed B and U, γ decreases with the increase of m. As a result, if we repeat a particular MI for n times, the mean of the γ_m would not converge to γ_0 when n goes to infinite. The relationship between γ_m and γ_0 is not the same as that between the sample mean $\bar{x}$ and the population mean μ. The expected value of $\bar{x}$ is μ, whereas the expected value of γ_m is not γ_0. For a finite m, the expected value of γ_m is always greater than the corresponding γ_0. In the literature, researchers may have used γ_m or the means of γ_m of a relatively small m (≤ 20) as an estimate of γ_0. As a result, the γ values presented in some literatures may be much inflated.

Let B_0 and U_0 be the expected value of B and U and the total variance T_0 be the sum of B_0 and U_0, i.e. $T_0 = B_0 + U_0$. By definition [4], γ_0 is the ratio of B_0 to T_0, i.e. $\gamma_0 = B_0/T_0$. From equations (2) , (3) and (4), we can prove that $\gamma_\infty = B_0/T_0 = \gamma_0$. Since for the same data, B_0, U_0 and T_0 would be affected by how the data are analysed, γ_0 would not be unique for the same data. The same dataset may have many different γ_0 values. In Figures 1 and 2, the γ values was the mean of γ_m of 30 replicates. The level off of the m-γ curve when m approached 99 indicate that m=99 in these MI trials of this study can be regarded as an approximation of m=∞, and the γ values at m=99 can be regarded as the γ_0. For the same ImpV and δ at m=99, the difference in γ values between different AnaV variables experimentally proved that that same data may have different γ_0 values (Figures 1 and 2), which shakes the foundation of using γ_0-based method to determine the sufficient m.

Van Buuren (2012) pointed out that the scope of MI can be broad, intermediate, or narrow [1]. For MI of broad and intermediate scopes, the complete datasets generated by MI may be analysed in multiple ways. Government agencies such as NCHS conduct national surveys and release data to the public. The MI for the data of these national surveys such as PWS should be targeted for analyses in many ways by various data users. Therefore, the same survey data will definitely have many γ_0 values. Use of γ_0 to determine sufficient numbers of m for these surveys may not be appropriate.

5. Conclusions

For $m < \infty$, the γ value decreases with the increase of m. As a result, the expected value of γ_m does not converge to γ_0. The relationship between γ_m and γ_0 is not the same as that between the sample mean $\bar{x}$ and the population mean μ. Unlike δ, which is a feature of the data and will not change with how the data are analysed, γ is not a feature of the data and may change greatly with the ways the data are analysed. There may be many γ_0 values for the same data. As a results, the method of determining the sufficient m based on the γ_0-based RE is lack of solid foundation.

References

[1] S. Van Buuren, 2012, Flexible Imputation of Missing Data, Chapter 2. Multiple imputation. Boca Raton, FL: Chapman and Hall / CRC Press, pp. 25-52.

[2] P.H. Rezvan, K.J. Lee and J.A. Simpson, 2015, "The rise of multiple imputation: a review of the reporting and implementation of the method in medical research," BMC Medical Research Methodology 15: 30.

[3] Y. Deng, C. Chang, M.S. Ido, and Q. Long, 2016, "Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data," Scientific Reports 6:21689, DOI: 10.1038/srep21689.

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

[5] Q. Pan and R. Wei, 2016, "Fraction of missing information (γ) at different missing data fractions in the 2012 NAMCS Physician Workflow Mail Survey," Applied Mathematics, 2016 (7), pp. 1057-1067.

[6] Q. Pan and R. Wei, 2015. "Relationship between missing information and missing data in 2012 NAMCS Physician Workflow Mail Survey," In 2015 JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association. pp. 2630-2637.

[7] Q. Pan, R. Wei, I. Shimizu and E. Jamoom, 2014, "Determining Sufficient Number of Imputations Using Variance of Imputation Variances: Data from 2012 NAMCS Physician Workflow Mail Survey," Applied Mathematics, 2014, 5, 3421-3430.

[8] Q. Pan, R. Wei, I. Shimizu and E. Jamoom, 2014. "Variances of Imputation Variances as Determiner of Sufficient Number of Imputations Using data from 2012 NAMCS Physician Workflow Mail Survey," In 2014 JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association. 3276-3283.

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

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

[11] J.B. Asendorpf et al., 2014, "Reducing bias due to systematic attrition in longitudinal studies: The benefits of multiple imputation," International Journal of Behavioral Development **38**(5): 453-460.

[12] J.W.Bartlett, et al., 2015, "Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model," Statistical Methods in Medical Research 24(4): 462-487.

[13] X.Basagana, et al. 2013, "A framework for multiple imputation in cluster analysis," American Journal of Epidemiology 177(7): 718-725.

[14] K.Biering, et al., 2015, "Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes", Clinical Epidemiology 7: 91-106.

[15] Lau, D.T., McCaig, L.F., and Hing, E. (2016) "Toward a More Complete Picture of Outpatient, Office-Based Health Care in the U.S.: Expansion of NAMCS", Am. J. Prev. Med. 2016 Apr 5. pii: S0749-3797(16)30003-4. doi: 10.1016/j.amepre.2016.02.028. [Epub ahead of print]

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

[17] R.R. Andridge and R.J.A. Little, 2010, "A review of hot deck imputation for survey non-response", Int Stat Rev 78(1): 40–64.