Relationship between missing information and missing data in 2012 NAMCS Physician Workflow Mail Survey [1]

Qiyuan Pan*, Rong Wei
National Center for Health Statistics, Centers for Disease Control and Prevention, USA
*Corresponding author: qpan@cdc.gov

Abstract

The fraction of missing information, γ, is an important concept in multiple imputation (MI). Rubin (1987) used γ to define the relative efficiency (RE) of MI as $RE = (1+\gamma/m)^{-1/2}$, where m is the number of imputations, leading to the conclusion that only a small m, e.g. m = 2 or 3, would be sufficient where γ≤0.5. However, increasing evidence has shown that many more imputations are needed. Why would the apparently sufficient m deduced from the RE be actually too small? The answer may lie with the characteristics of γ, the only factor other than m itself that defines the RE. To date little research on γ has been done using real survey data. The relationship between γ and the fractions of missing data (δ) was studied using the 2012 National Ambulatory Medical Care Survey (NAMCS) Physician Workflow Mail Survey. The results suggest that γ and δ are not comparable, that it is impossible to predict γ using δ, and that the γ-based RE may be inappropriate in determining sufficient m.

Key Words: Multiple imputation, fraction of missing information (γ), sufficient number of imputations, missing data, NAMCS

1. Introduction

The importance of γ was signified when Rubin, in his classic book "Multiple Imputation for Nonresponse in Surveys" published in 1987, used γ to define the relative efficiency of MI (RE) as [1]:

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

Based on this RE, Rubin drew the following conclusion: If $\gamma \leq 0.2$, even two repeated imputations appear to result in accurate levels, and three repeated imputations result in accurate levels even when $\gamma = 0.5$ [1]. Rubin's conclusion of small m as being sufficient have been having huge impact on MI applications [2, 3].

For a limited number of imputations in MI, γ is estimated by the following equation [1]:

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

where r is the relative increase in variance due to nonresponse and v is the degrees of freedom, defined by equations (3) and (4) below, respectively [1]:

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

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

[1] The views of this paper do not necessarily reflect the views of the National Center for Health Statistics (NCHS) or the Centers for Disease Control and Prevention (CDC) of the United States government.

where B is the between-imputation variance and U is the within-imputation variance, defined by equations (5) and (6) below, respectively [1]:

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

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

where the subscript i denotes the *i*th imputation and Q is the quantity of interest. As m approaches infinity, the following relationship can be deduced from equations (2) to (6):

$$\gamma_{m\to\infty} = \frac{B_{m\to\infty}}{B_{m\to\infty}+U_{m\to\infty}} = \frac{B_{m\to\infty}}{T_{m\to\infty}}, \quad (7)$$

where $T_{m\to\infty}$ is the total variance (T) as m approaches infinity. Equation (7) shows that γ is ultimately the fraction of B in T, and it is termed as "the fraction of missing information" probably because B would be otherwise missing from T unless MI is used [1].

In recent years, however, there is undeniable evidence that a much greater number of imputations, e.g. 40 or more, are needed in order to obtain reliable statistical inferences [4, 5, 6, 7, 8, 9, 10, 11]. On the one hand, statistical software packages such as SPSS and SAS still uses m=5 as the default value for the MI procedure, showing the persisting impact of Rubin's recommendation of small m as being sufficient. On the other hand, most researchers have realized that m≤5 is too small and are now using 40 or more imputations in their MI applications [12, 13, 14, 15].

Why would the apparently sufficient m as suggested by the γ-based RE be actually insufficient? The answer may lie with the characteristics of γ, for it is the only factor that defines the RE other than m itself. To date very little is known about the characteristics of γ other than what was described by Rubin in 1987 [1]. Even though many surveys are using MI, no published literatures can be found showing that γ values are determined using real survey data prior to the selection of sufficient m.

Fraction of missing information sounds similar to fraction of missing data (δ). In fact, "data" and "information" can largely be considered synonyms. How are γ and δ related? Rubin stated that γ would be equal to the expected δ in the simple case of no covariates, and commonly less than δ when there are covariates [1]. Using the 2012 Physician Workflow Mail Survey (PWS12) of the National Ambulatory Medical Care Survey (NAMCS), the relationship between γ and δ is examined. The data presented in this paper add to our understanding of γ and help explain why Rubin's γ-based conclusion on sufficient m may actually be too small.

2. Methodology

Conducted by the National Center for Health Statistics (NCHS), the NAMCS Physician Workflow Mail Survey (PWS) was a nationally representative, 3-year (2011-2013) panel mail survey of office-based physicians, with each year being a complete survey cycle [16]. The data of the 2012 PWS, i.e. PWS12, were used in this research. PWS12 had 2,567 eligible, responding physicians in the sample. Three variables representing the physician's practice size (SIZE), namely SIZE5, SIZE20 and SIZE100, were selected as the variables for imputation. SIZE100 is the practice size as represented by the number of physicians ranging from 1 to 100. SIZE5 and SIZE20 were derived from SIZE100. SIZE5 was derived by recoding the values of SIZE100 into 5 categories, and SIZE20 was derived by top-coding the values of SIZE100 greater than 20 into 20. These three variables differed in their value ranges, distributions, and variances (Table 1).

Table 1: Characteristics of the imputation variables.

Variable	Description	Mean	Value range	Total variance
SIZE5	Practice size recoded from SIZE100: 1 = Solo practice; 2 = Two physicians; 3 = 3 to 5 physicians; 4 = 6-10 physicians; 5 = 11+ physicians.	3.06	1 – 5	1.97
SIZE20	Practice size recoded from SIZE100: 1-19 = The actual number of physicians; 20 = 20+ physicians.	6.47	1 – 20	38.26
SIZE100	Practice size as represented by the number of physicians.	11.41	1 – 100	483.02

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 non-missing values from the 2011 data for the same physician, the δ of PWS12 became 4%. The two other two δ values, 10% and 20%, were obtained by partially replacing, in a random manner, the missing values in 2012 with the non-missing values in 2011 survey for the same physician. This method assumes that the value of SIZE would not change for the same physicians between 2011 and 2012. The method was officially used by NCHS in producing the public use data from PWS12. Therefore the δ values 4%, 10%, and 20% may be considered as the survey data instead of simulation data.

Hot deck imputation [17] was used. The statistics software package SAS 9.3 was used to carry out the imputation procedure. For each imputation variable at each δ, 1,000 independent imputations were done. From this pool of 1,000 imputations, samples were randomly drawn to form MI of various m values. For the purpose of this study, m=80 was chosen. To calculate the variance of γ, 30 random MI samples of m=80 were drawn. Four analytic treatments were used (Table 2). Analyses were conducted with the un-weighted data.

3. Results and discussions

3.1 The γ-δ relationship for different analytic treatments

The shapes of the lines showing the changes of γ as affected by changes in δ were drastically different among the four analytic treatments, indicating that the choice of analytic variables had major effect on γ, with γ varying from <0.0001 for CONTROL to >0.004 for DERIVED (Figure 1). The γ usually increased with the increase of the δ (Figure 1). However there are exceptions. For PRIMEMP, γ did not increase when δ increased from 20% to 29%. For DERIVED, γ decreased sharply when δ increased from 4% to 10% (Figure 1 a).

3.2 The γ-δ relationship for different categories of an analytic variable

Data of the first four categories in the value list of PRIMEMP (Table 2) are presented in Figure 2 as examples to show whether different values of an analytic variable may have different γ values and γ-δ relationships. The four PRIMEMP categories had different γ values and γ-δ relationships, as visualized by the line graphs of Figure 2. When δ increased, γ usually increased but could decrease or remain unchanged (Figure 2).

3.3 Effects of imputation variables on γ-δ relationship

As shown in Figure 3, SIZE5, SIZE20 and SIZE100 had similar γ values at δ=4%. As δ increased, there was an increase in the differences of their γ values among the three

imputation variables. Usually the greater the variance of the imputation variable, the greater the γ value (compare Table 1 and Figure 3). For all three imputation variables, γ increased when δ increased from 4% to 20%, then stabilized when δ further increased from 20% to 29% (Figure 3).

Table 2: Description of the analytic treatments

Treatment	Description	Value range	Correlation coefficient[1]		
			SIZE5	SIZE20	SIZE100
CONTROL	No analytic variable	N/A[2]	N/A[2]	N/A[2]	N/A[2]
REGION	Region of the Physicians Interview office	1=Northeast, 2=Mid West, 3=South, 4=West	-0.0766	-0.0264	-0.0551
PRIMEMP	Primary present employment of the physician	11=AMA-Self-emp, solo prac; 13=AMA-Two phy. prac; 20=AOA-Office prac. solo; 21=AMA-Oth pat care/AOA-Off prac. partnp; 22=AOA-Office prac group; 23=AOA-Offc prac ofc employee; 30=AMA-Grp prac/AOA-Off prac HMO staff; 31=AOA-Office prac. walk-in clinic; 35=AMA-HMO; 40=AMA-Medical school; 64=AMA-County/Cty/State Govt Other; 97=AOA-other office or clinic practice; 110=AMA-No classification; 200=Sampled CHC	0.2180	0.0554	0.1341
DERIVED	Derived categories for SIZE5, SIZE20, and SIZE100	Regrouping the values with random errors added to each group. 1 to 4 for SIZE5, 1 to 9 for SIZE20, 1 to 17 for SIZE 100	0.9659	0.9954	0.9109

[1]The correlation coefficient between the practice size and the values of analytic treatments.
[2]N/A: Not applicable.

3.4 The magnitude of γ

The mean γ values for the four analytic treatments varied from 0.000043 to 0.0059 (Table 3). With such a small γ, Rubin's γ-based RE would become 1 even at m=1, which could be interpreted as that single imputation was sufficient and MI might be meaningless. The data suggest that Rubin's γ-based RE may not be appropriate to determine the sufficient m for MI.

Table 3: γ values of different analytic treatments

Analytic treatments	Mean	Minimum	Maximum
CONTROL	0.000043	0.0000087	0.000097
REGION	0.000241	0.0000478	0.000500
PRIMEMP	0.003943	0.0003935	0.007911
DERIVED	0.005945	0.0000438	0.017999

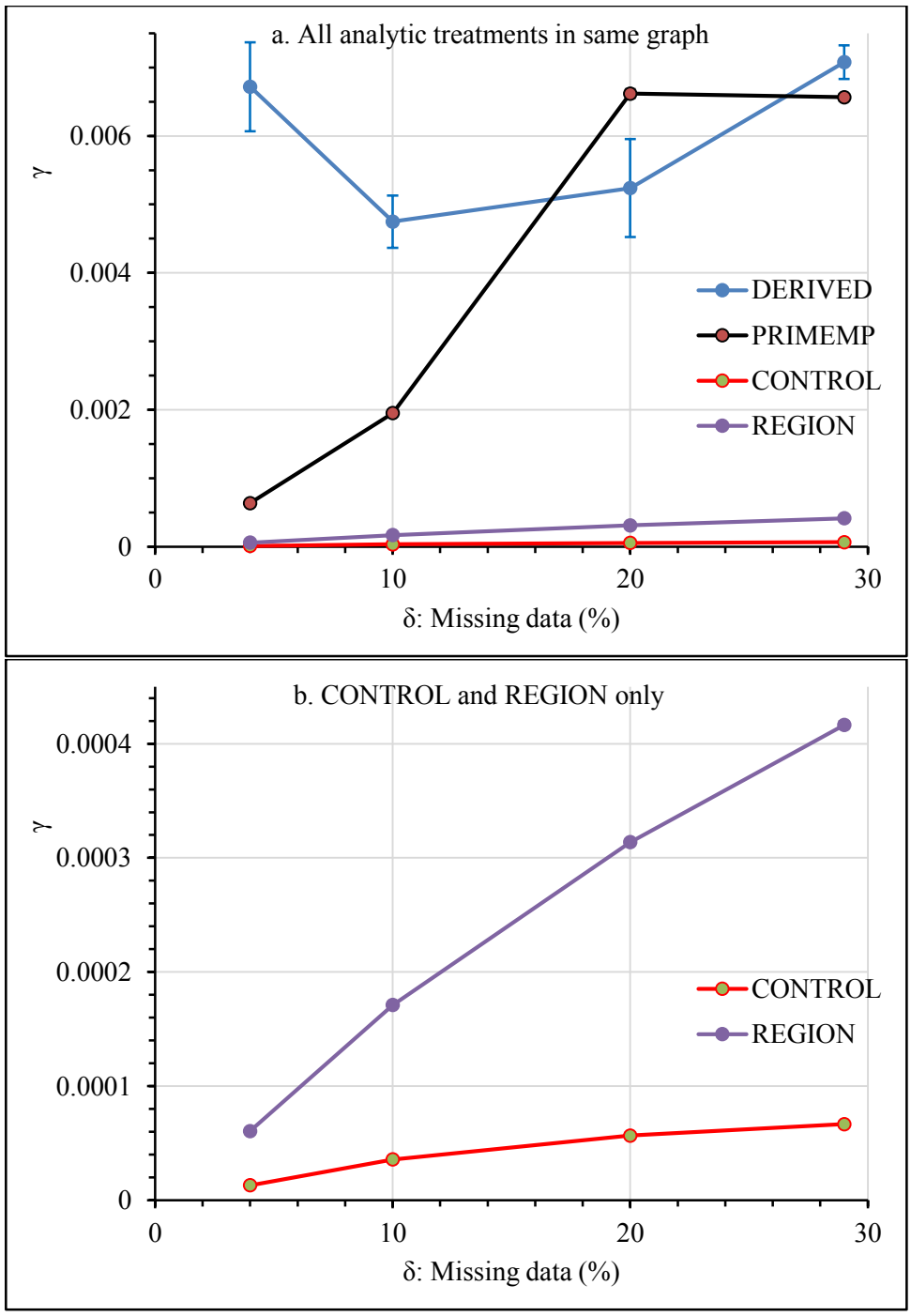

Figure 1: Effects of the analytic treatments on the γ-δ relationship: a. All the four analytic treatments are presented in the same graph; b. CONTROL and REGION, which had much smaller γ than PRIMEMP and DERIVED, are graphed separately so that their γ-δ relationship can be better visualized. Note: The standard errors for all data points of CONTROL, REGION, and PRIMEMP were too small to be visualized in the graph.

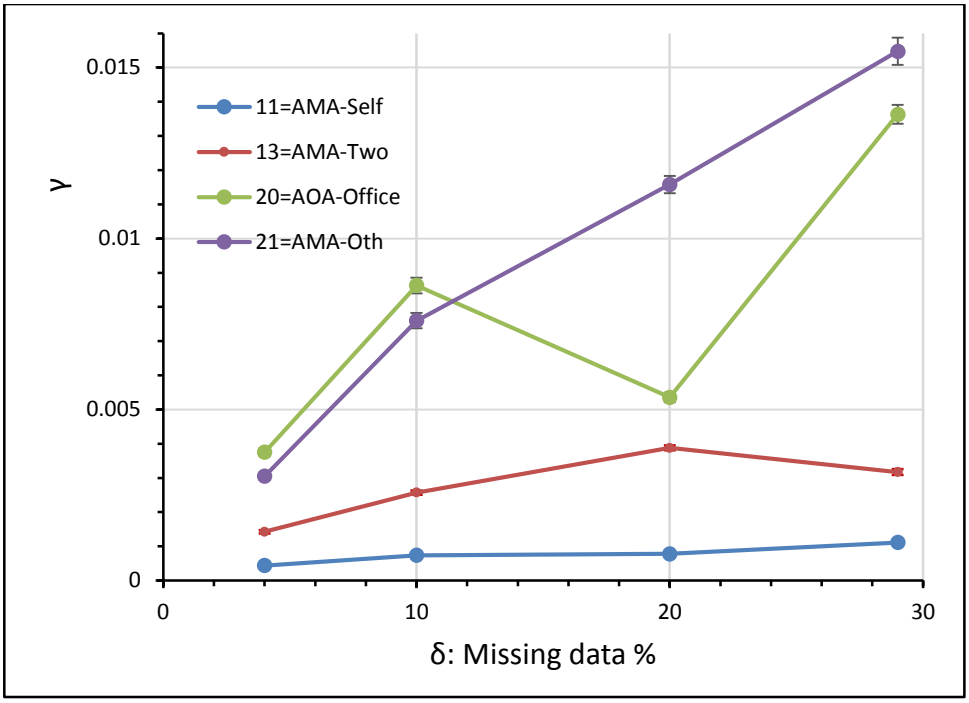

Figure 2: Effects of different categories of PRIMEMP on the γ-δ relationship. Data of the first four categories of PRIMEMP (Table 2) are presented.

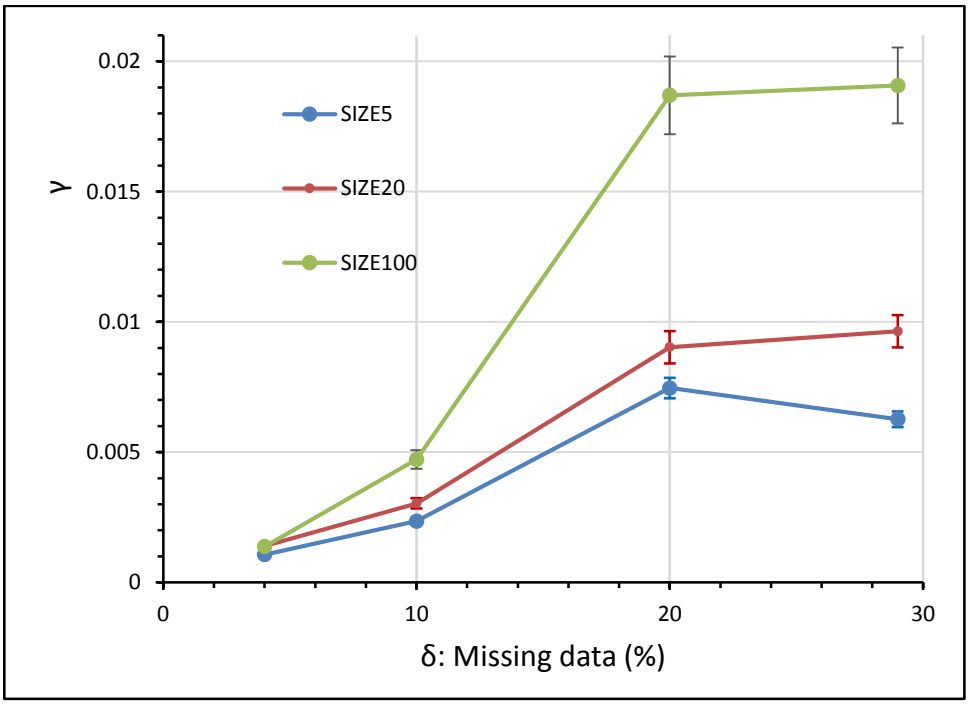

Figure 3: Effects of imputation variables on the γ-δ relationship.

3.5 Additional discussions

For PWS12, γ is one, two, or sometimes three orders of magnitude less than δ. This enormous difference between γ and δ cannot be possibly explained by the existence of covariates as suggested by Rubin [1]. γ is essentially a ratio of variances whose value

can be affected by any factors that affect the variance and partitioning of the variance among B and U, whereas δ is simply a ratio of sampling unit counts whose value is fixed once the survey is done.

With $m \to \infty$, γ becomes B/T, where T=B+U (equation (7)). For the MI experiments in this study, sample size is n=2,567 for CONTROL, the Q for equation (6) is the sample mean, B is the variance of the sample means in nature, and U is the mean of the sample variances. Based on the classic statistics, the sample variance s^2 and the variance of the sample means $s_{\bar{x}}^2$ is $s_{\bar{x}}^2 = s^2/n$ [18]. Therefore, B can be estimated by U/n = U/2567, and γ can be estimated by B/T = U/(2567T), which would be ≤1/2567 for the CONTROL. Factors such as how large the δ is and whether covariates exist may affect how much γ is smaller than 1/2567 or about 0.00039, but cannot, theoretically, make γ bigger than 1/2567. This may explain why γ was so small in this study. This may also imply that the mathematical base for γ=δ may not exist even in the simple case of no covariates.

The CONTROL treatment in this study assumes no covariate. Under the situation of CONTROL, γ linearly increased with δ (Figure 1 b). But the γ values are so much smaller than δ. The γ values were also affected by factors such as imputation variables (Figure 3) and sample size (see discussions above). Therefore, it is impossible to predict γ using δ even if a linear relationship may exist between γ and δ under certain circumstances.

4. Conclusions

Rubin stated that γ would be equal to the expected δ in the simple case of no covariates, and commonly less than δ when there are covariates [1]. Results of PWF12 in this study suggest that γ and δ are drastically different and not comparable. The magnitude of γ in this study varied from 0.01 to 0.000001. The enormous difference between γ and δ cannot possibly be explained by the presence or absence of covariates. The supposition that γ = E[δ] is untenable. It is impossible to predict γ using δ even if a linear relationship may exist between them under certain circumstances. If the γ-based RE is used to determine the sufficient number of multiple imputations, for γ≤0.01, a single imputation would be sufficient and MI would become meaningless. An alternative interpretation for this result is that it may be inappropriate to use the γ-based RE to determine the sufficient number of m. Any factors affecting the variance and the partition of the variance between B and U would affect γ and the γ-δ relationship.

Acknowledgement

The authors sincerely thank Dr. Alan H. Dorfman, Office of Research and Methodology (ORM), NCHS, CDC, USA, for his valuable suggestions on the research and critical text editing of the paper.

References

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

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

[3] J.L. Schafer, 1997, Analysis of Incomplete Multivariate Data, Washington D.C.: Chapman and Hall / CRC, pp. 89-145.

[4] 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.

[5] 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.

[6] 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.

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

[8] J. L. Schafer and J. W. Graham, 2002, "Missing data: Our view of the state of the art", Psychological Methods, Vol. 7, pp. 147–177.

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

[10] P. Royston, 2004, "Multiple imputation of missing values," The Stata Journal, Vol. 4, No. 3, pp. 227-241.

[11] T.E. Bodner, 2008, "What improves with increased missing data imputations?" Structural Equation Modeling: A Multidisciplinary Journal 15: 651-675.

[12] 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.

[13] 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.

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

[15] 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.

[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.

[18] M.L. Berenson and D.M. Levine, 1990, Statistics for Business & Economics, Prentice Hall, pp 264-285.