HHS Public Access

Author manuscript

Cogent Math Stat. Author manuscript; available in PMC 2019 March 19.

Published in final edited form as:

Cogent Math Stat. 2018 ; 5: . doi:10.1080/25742558.2018.1551504.

Improved methods for estimating fraction of missing information in multiple imputation

Qiyuan Pan[1,*] and **Rong Wei**[2]

[1]National Center for Health Statistics (NCHS), 3311 Toledo Rd., Hyattsville, Maryland 20782, USA.

[2]NCHS, 3311 Toledo Rd., Hyattsville, MD 20782, USA.

Abstract

Multiple imputation (MI) has become the most popular approach in handling missing data. Closely associated with MI, the fraction of missing information (FMI) is an important parameter for diagnosing the impact of missing data. Currently γ_m, the sample value of FMI estimated from MI of a limited m, is used as the estimate of γ_0, the population value of FMI, where m is the number of imputations of the MI. This FMI estimation method, however, has never been adequately justified and evaluated. In this paper, we quantitatively demonstrated that $E(\gamma_m)$ decreases with the increase of m so that $E(\gamma_m) > \gamma_0$ for any finite m. As a result γ_m would inevitably overestimate γ_0. Three improved FMI estimation methods were proposed. The major conclusions were substantiated by the results of the MI trials using the data of the 2012 Physician Workflow Mail Survey of the National Ambulatory Medical Care Survey, USA.

Keywords

fraction of missing information; multiple imputation; missing data; National Ambulatory Medical Care Survey; number of imputations; Science; Mathematics & Statistics; Applied Mathematics; Mathematics for Biology & Medicine

[*]Corresponding author: Qiyuan Pan, National Center for Health Statistics, USA qap1@cdc.gov.

ABOUT THE AUTHORS

Qiyuan Pan is a mathematical statistician in Division of Health Care Statistics, National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), U.S. Department of Health and Human Services (HHS), USA. He obtained his PhD from University of Maryland, USA. Rong Wei is a mathematical statistician in Mathematical Statistics Branch, Division of Research and Methodology, NCHS, CDC, HHS. She obtained her PhD from University of Wisconsin, USA.

PUBLIC INTEREST STATEMENT

In any big surveys such as the National Ambulatory Medical Care Survey, USA, you select people into your sample and ask them to answer your questions. Some will answer, and others will not answer. When you finally get your survey data, it is inevitable that you will have missing data due to those non-respondents. How to minimize the effects of the missing data on your survey has been a big topic in statistics. Multiple imputation (MI) has become the most popular approach in handling missing data. Closely associated with MI, the fraction of missing information (FMI) is an important parameter for diagnosing the impact of missing data. This paper shows that the current method for estimating FMI bears intolerable biases. We proposed three improved methods that would give more accurate estimate of FMI.

Competing interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

1. Introduction

Multiple imputation (MI) becomes the most popular approach to accounting for missing data (Carpenter & Kenward, 2013, Dohoo, 2015, Rezvan, Lee, & Simpson, 2015, Rubin, 1987, Van Buuren, 2012). Closely associated with MI, fraction of missing information (FMI) is an important parameter for diagnosing the effects of data missingness (Rubin, 1987). FMI can be interpreted as the fraction of information about Q due to non-response, where Q is the quantity of interest (Rubin, 1987). As MI become increasingly important, the importance of FMI is also increasing. The best known use of FMI is to define the relative efficiency (RE) of MI as RE = $(1 + \gamma_0/m)^{-1/2}$, where γ_0 is the population value of FMI and m is the number of imputations (Rubin, 1987). Based on this RE, Rubin concluded that $m \leq 5$ would be sufficient for MI (Rubin, 1987). Little et al. as well as Wagner suggested that FMI be used as an alternative tool for measuring data missing data or the response rate (Little et al., 2016, Wagner, 2010). Siddique, Harel, Crespic, and Hedekerd (2014) used FMI to verify the missing data mechanisms. The most common practice of FMI estimation is to use $\hat{\gamma}_0 = \gamma_m$, where $\hat{\gamma}_0$ is the estimated value of γ_0 and γ_m is the FMI obtained from MI of a given m, e.g. (Khare, Little, Rubin, & Schafer, 1993, Lewis et al., 2014, Schafer, 2001, Schenker et al., 2006). However, the accuracy of the $\hat{\gamma}_0 = \gamma_m$ method has not been adequately evaluated. This paper is to quantify possible biases of $\hat{\gamma}_0 = \gamma_m$ and to improve FMI estimation methodology if necessary and possible.

Established by Rubin in 1987, the current FMI paradigm is defined by Equations (1)–(11) below:

$$\bar{Q}_m = \frac{1}{m}\sum_{m}^{1} Q_i, \quad (1)$$

where subscript m and ∞ stands for a finite and infinite m, the subscript 0 for the population value, the subscript i for the ith imputation, and the bar hat for the parameter's mean.

$$B_m = \frac{1}{m-1}\sum_{1}^{m}(Q_i - \bar{Q}_m)^2 \quad (2)$$

$$U_m = \frac{1}{m}\sum_{1}^{m} U_i \quad (3)$$

$$T_m = U_m + \left(1 + \frac{1}{m}\right)B_m, \quad (4)$$

where B, U, and T are the between-imputation, within-imputation, and the total variances.

$$r = \left(1 + \frac{1}{m}\right)\frac{B_m}{U_m}, \quad (5)$$

where r is the fractional variance increase due to data missingness.

$$v = (m-1)\left(1 + \frac{1}{r}\right)^2, \quad (6)$$

where v is the degrees of freedom.

$$\gamma_m = \frac{r + 2/(v+3)}{r+1} \quad (7)$$

$$T_\infty = U_\infty + B_\infty \quad (8)$$

$$\gamma_\infty = \frac{B_\infty}{T_\infty} \quad (9)$$

$$T_0 = U_0 + B_0 \quad (10)$$

$$\gamma_0 = \frac{B_0}{T_0}. \quad (11)$$

Equation (11) cannot be used to calculate γ_0 in practice because B_0, U_0, and T_0 are usually unknown. No researchers have provided an equation that explicitly links γ_m and γ_0. The justification for using $\hat{\gamma}_0 = \gamma_m$ is not available from Equations (1)–(11).

Assume $\gamma_\infty = \gamma_0$. For $\hat{\gamma}_0 = \gamma_m$ to be valid, $E(\gamma_m) = \gamma_0$ must be true. For $E(\gamma_m) = \gamma_0$ to be true, $E(\gamma_m)$ must be independent of m. To understand $E(\gamma_m)$, let the same MI of a given m be repeated for j times. Denote the γ_m from each MI repeat as γ_{m1}, γ_{m2}, γ_{m3}, …, γ_{mj}. By definition, the expected value is the sum of all possible values each multiplied by the probability of its occurrence (Hogg, McKean, & Craig, 2013). Therefore, $E(\gamma_m)$ can be defined as:

$$E(\gamma_m) = \lim_{j \to \infty} \left(\frac{1}{j} \sum_{j}^{1} \gamma_{mj} \right). \quad (12)$$

Equation (12) shows that $E(\gamma_m)$ can be understood as the ultimate mean of γ_m when j becomes infinity. If $E(\gamma_m)$ is independent of m, we should have $E(\gamma_2) = E(\gamma_3) = \ldots = E(\gamma_m) = \gamma_0$, and the use of $\hat{\gamma}_0 = \gamma_m$ would be justified. If $E(\gamma_m)$ depends on m, we should have $E(\gamma_2) \neq E(\gamma_3) \neq \ldots \neq E(\gamma_m) \neq \gamma_0$. The use of $\hat{\gamma}_0 = \gamma_m$ may not be justified if the difference between $E(\gamma_m)$ and γ_0 is intolerably big.

Rubin indicated that the mean of γ_m can be regarded as γ_0 [1 page 143], underlining an assumption that $E(\gamma_m)$ is independent of m. Harel briefly mentioned that γ_m "tends to decrease as m increases" without providing any details (Harel, 2007). Although Harel's statement favours $E(\gamma_m) \neq \gamma_0$, it cannot be a base for disproving $\hat{\gamma}_0 = \gamma_m$ because it might be acceptable to use $\hat{\gamma}_0 = \gamma_m$ if the decrease of γ_m with the increase of m is statistically negligible. To date the justifications for using $\hat{\gamma}_0 = \gamma_m$ is still missing.

For FMI estimation, Harel's 2007 paper (Harel, 2007) is important in that it pointed out that, unlike γ_m that "tends to decrease as m increases," the quantity $B_m/(U_m + B_m)$ "does not tend to decrease as m increases" (Harel, 2007). Harel used $\hat{\gamma}_0 = B_m/(U_m + B_m)$ to estimate FMI in his research (Harel, 2007). The goal of Harel's paper was not to find a better FMI estimation method per se and his discussion on $\hat{\gamma}_0 = B_m/(U_m + B_m)$ was brief. Most researchers have not used Harel's method for FMI estimation probably because most people may have treated Harel's method as being research-specific rather than a method that may potentially be universally used for FMI estimation.

In this study, we examined the relationships between m, γ_m, γ_∞, and γ_0, quantified the decrease of γ_m with the increase of m, quantified the biases of $\hat{\gamma}_0 = \gamma_m$, and proposed improved methods for FMI estimation. Only univariate FMI definition will be examined in this paper even though multi-variate FMI definition may exist. The major conclusions were substantiated by the MI trials using the data of the 2012 Physician Workflow Mail Survey (PWS) of the National Ambulatory Medical Care Survey (NAMCS). This paper focuses on MI approach only even though it may be possible to estimate FMI via a non-MI approach (Savalei & Rhemtulla, 2012, Zheng & Lo, 2008).

2. The relationships between m, γ_m, γ_∞, and γ_0

2.1. The condition for $\gamma_\infty = \gamma_0$

To use γ_m for estimating γ_0, one must assume $\gamma_\infty = \gamma_0$. Most researchers simply treat γ_∞ and γ_0 as synonyms (e.g. He et al., 2016). But they are not. Imagine a population of imputations (POI) is generated by repeating the imputation of the same model on the same data for an infinite number of times. An MI is simply a sample of the POI with sample size

m. The sample value and the population value of FMI for the POI are γ_m and γ_∞, respectively. The population for γ_0, however, is not POI but the population of the sampling units of the survey. A γ_∞ is inseparably linked to an MI, but a γ_0 can be independent of MI. The γ_0 may be estimated by MI as well as other methods such as maximum likelihood (Savalei & Rhemtulla, 2012, Zheng & Lo, 2008). Using Equations (5)–(7), one can prove that the condition for $\gamma_\infty = \gamma_0$ is $B_m/U_m = B_0/U_0$. When we use MI to estimate γ_0, we have to assume $\gamma_\infty = \gamma_0$, which is probably why γ_∞ and γ_0 are often treated as synonyms in MI analyses. In this paper, we will assume $\gamma_\infty = \gamma_0$ because we use MI to estimate γ_0.

2.2. B_m/U_m is independent of m

Equation (7) that defines γ_m does not have m as a factor. Combining Equations (5), (6), and (7) gives an expanded definition of γ_m with m as one of the independent factors affecting γ_m:

$$\gamma_m = \frac{\left(1+\frac{1}{m}\right)\frac{B_m}{U_m} + 2/\left((m-1)\left(1+\frac{1}{\left(1+\frac{1}{m}\right)\frac{B_m}{U_m}}\right)^2 + 3\right)}{\left(1+\frac{1}{m}\right)\frac{B_m}{U_m}+1}. \quad (13)$$

Equation (13) shows that γ_m is a function of three factors, i.e. $\gamma_m = F(m, B_m, U_m)$. In Equation (13), B_m and U_m always appear together as B_m/U_m. Letting $c_m = B_m/U_m$, then γ_m becomes a function of two factors, i.e. $\gamma_m = F(m, c_m)$.

Whether c_m is independent of m is important in understanding the m–γ_m relationship. If c_m depends on m, the direct effects of m on γ_m would be confounded by the indirect effects of m on γ_m via m's effects on c_m, which in turn could be due to m's effect on B_m, U_m or both. If c_m is independent of m, then the m–γ_m relationship would be greatly simplified.

In order to establish that c_m is independent of m, we need to prove $E(B_m/U_m) = B_0/U_0$. Equation (2) indicates that the relationship between m and B_m is that between the sample size n and the variance (s^2) so that $E(B_m)$ is independent of m, i.e. $E(B_m) = B_0$ (Serfling, 1980). Equation (3) indicates that the relationship between m and U_m is that between the sample size n and the sample mean $\bar{x}$ so that $E(U_m)$ is independent of m, i.e. $E(U_m) = U_0$ (Hogg et al., 2013). Jensen's Inequality (Hogg et al., 2013) determines that $E(1/U_m) \geq 1/E(U_m)$. Therefore $E(B_m/U_m) = E(B_m)E(1/U_m) \geq E(B_m)/E(U_m) = B_0/U_0$, or $E(B_m/U_m) \geq B_0/U_0$. Our simulation studies show that the maximum difference between $E(B_m)/E(U_m)$ and $E(B_m/U_m)$ is less than 0.1%, which is negligible in virtually any statistics work. We can safely regard $E(B_m/U_m) = B_0/U_0$ as a fact for the purpose of studying the m–γ_m relationship. The c_m's independence of m is thus proved. The subscript m can be removed from c_m. As a result, we can indeed letting $c_m = B_m/U_m$ be a constant c in Equation (13) and make γ_m become a function of the single factor m, i.e. $\gamma_m = F(m)$.

2.3. The $\gamma_m = F(m,\gamma_0)$ equation

When m goes infinite, γ_m becomes γ_0. Our goal is to establish the mathematic relationship between γ_m and γ_0 at a finite m, which is currently missing in published literatures. In the discussions above, we have showed that it is mathematically legitimate to letting c_m be a constant c in studying the m– γ_m relationship because c_m is independent of m. What is the best value to choose for c to obtain the most truthful m–γ_m relationship? The answer is: $c = E(B_m/U_m) = B_0/U_0$. If and only if $c = E(B_m/U_m) = B_0/U_0$, the m–γ_m relationship as determined by Equation (13) would reflect the true m–γ_m relationship. From Equations (10) and (11) we can obtain $B_0/U_0 = \gamma_0/(1-\gamma_0)$. Replacing B_m/U_m in Equation (13) with $\gamma_0/(1-\gamma_0)$, we obtain an equation that directly links γ_m to γ_0 as follows:

$$\gamma_m = E(\gamma_m) = F(m,\gamma_0) = \frac{\left(1+\frac{1}{m}\right)\frac{\gamma_0}{1-\gamma_0} + 2/\left((m-1)\left(1+\frac{1}{\left(1+\frac{1}{m}\right)\frac{\gamma_0}{1-\gamma_0}}\right)^2 + 3\right)}{\left(1+\frac{1}{m}\right)\frac{\gamma_0}{1-\gamma_0}+1}. \quad (14)$$

Establishment of equation is a significant step forward in understanding the relationship between m, γ_m, and γ_0 because it links the three factors in the same equation for any m, finite or infinite.

For a given analysis of a given dataset, γ_0 is a constant. When we repeat the same MI of a given m for j times, the γ_m value from each repeat of the MI will not change when the γ_m is determined by Equation (14). In other words, we will have $\gamma_{m1} = \gamma_{m2} = \gamma_{m3} = \cdots \gamma_{mj} = E(\gamma_m)$ (see Equation (12)). In other words, the γ_m value obtained from Equation (14) will be $E(\gamma_m)$. For different data and analyses, γ_0 is a variable. Equation (14) shows that $E(\gamma_m)$ is a function of the two factors, m and γ_0, i.e. $E(\gamma_m) = F(m, \gamma_0)$.

3. The decrease of $E(\gamma_m)$ with the increase of m

3.1. $E(\gamma_m) > \gamma_0$ for any finite m

We all know that $E(\bar{x})$ is independent of n and equals to μ, which provides the theoretical base for $\hat{\mu} = \bar{x}$ (Hogg et al., 2013). The use of $\hat{\gamma}_0 = \gamma_m$ implies the assumption that $E(\gamma_m) = \gamma_0$. Using Equation (14), the m–$E(\gamma_m)$ relationship curve can be constructed for any given γ_0. Figure 1 presents the m–$E(\gamma_m)$ relationship curves for $\gamma_0 = 0.15$ and 0.2. Based on Figure 1, for the first time in MI research, we can explicitly state this important fact: $E(\gamma_m)$ decreases with the increase of m. The decrease of $E(\gamma_m)$ with the increase of m can be interpreted as follows: For a given dataset with a given MI model, the ultimate mean of γ_m, which is the mean of an infinite number of individual γ_m values obtained from repeating the MI of the given m for an infinite number of times, would always be greater than the γ_0. Of course what is called “the ultimate mean” here is the $E(\gamma_m)$ (Hogg et al., 2013). Therefore, by showing $E(\gamma_m)$ decreases with the increase of m, we have proved that $E(\gamma_m) > \gamma_0$ for any

finite m (Figure 1). The fact that $E(\gamma_m) > \gamma_0$ is further illustrated by more data in Table 1 for a wider range of m values and more γ_0 values.

3.2. The bias of the current FMI estimation method

The fact that $E(\gamma_m) > \gamma_0$ dictates that the current FMI estimation method $\hat{\gamma}_0 = \gamma_m$ must be biased. One achievement of this paper is that we successfully quantified the bias of the current FMI estimation method. We use D_γ, the percentage difference between $E(\gamma_m)$ and γ_0 as the parameter to measure this bias, i.e.:

$$D_\gamma = 100\frac{\gamma_m - \gamma_0}{\gamma_0}. \quad (15)$$

Table 1 presents the D_γ values at different γ_0 and m values as determined by Equation (14). At a given m, D_γ differs at different γ_0 values (Table 1). For $m = 2$, D_γ is 80.59% and 53.64% for $\gamma_0 = 0.2$ and 0.01, respectively (Table 1). When γ_0 increased from 0.001 to 0.6, D_γ first increases with the increase of γ_0, reaches a peak, and then decreases (Table 1 and Figure 2). The value of the γ_0 at which D_γ reaches the peak differs with m (data not shown). For $m = 5$, the maximum D_γ value of 25.31% occurs at $\gamma_0 = 0.23$, and the minimum D_γ value of 16.53% occurs at $\gamma_0 = 0.6$ (Figure 2(b)). In other words, one could overestimate FMI by 25% at $m = 5$ if the current method is used. A bias of this magnitude cannot and should not be ignored. Development of a better FMI estimation method is indeed necessary.

3.3. The γ_m decrease rate: smaller at larger m

We use $R_{\gamma d}$, the percentage rate of the γ_m decrease per unit m, to measure the rate of the γ_m decrease:

$$R_{D\gamma} = 100\frac{\gamma_m - \gamma_{m+1}}{\gamma_{m+1}}. \quad (16)$$

$R_{D\gamma}$ is affected by both m and γ_0 (Table 1 and Figure 2, b1 and b2). At $m = 5$, $R_{D\gamma}$ is 3.87% and 2.96% for $\gamma_0 = 0.2$ and 0.01, respectively (Table 1). Figure 2(b) show that $R_{D\gamma}$ increases initially, reaches a peak, and then decreases as γ_0 increases from 0.001 to 0.6. For $m = 2$, the maximum $R_{D\gamma} = 23.91\%$ occurs at $\gamma_0 = 0.15$ (Figure 2(b)). For $m = 5$, the maximum $R_{D\gamma} = 3.88\%$ occurs at $\gamma_0 = 0.21$ (data not shown). The gradual reduction of $R_{D\gamma}$ makes it possible for choosing a sufficient m when the m-driven γ_m reduction becomes negligibly small.

4. Improved methods for γ_0 estimation

Regarding $\hat{\gamma}_0 = \gamma_m$ as the control, any method that gives more accurate FMI estimation than this control will be considered as an improved method. Three improved methods are proposed below.

4.1. Improved method 1: $\hat{\gamma}_0 = \gamma_{m \geq 100}$

The control method is to use $\hat{\gamma}_0 = \gamma_m$ regardless the size of m. The first improved method is to choose a sufficiently large m when use $\hat{\gamma}_0 = \gamma_m$. Data in Figure 1 show that $E(\boldsymbol{\gamma}_m)$ approaches $\boldsymbol{\gamma}_0$ as m gets larger. Therefore, $\boldsymbol{\gamma}_m$ would estimate $\boldsymbol{\gamma}_0$ with an adequate accuracy when m is sufficiently large. Various criteria have been used to determine the sufficient m (Bodner, 2008, Graham, Olchowski, & Gilreath, 2007, Hershberger & Fisher, 2003, Pan, Wei, Shimizu, & Jamoom, 2014, Royston, 2004, Rubin, 1987). An adequately accurate estimation of $\boldsymbol{\gamma}_0$ using $\hat{\gamma}_0 = \gamma_m$ offers another criterion for determining a sufficient m. As measured by $R_{D\gamma}$, the gain in reducing the bias from increasing a unit m becomes smaller at a greater m. Using Equation (14), we can prove the bias of the default method as measured by $D_{\boldsymbol{\gamma}}$ would be about 1% or less for any reasonable $\boldsymbol{\gamma}_0$ values when m is greater than 100. We arbitrarily choose a bias of ≤1% as an acceptable level and recommend $m \geq 100$ as being sufficient for an adequately accurate estimation of $\boldsymbol{\gamma}_0$ using $\hat{\gamma}_0 = \gamma_m$. This method can be expressed as $\hat{\gamma}_0 = \gamma_{m \geq 100}$.

4.2. Improved method 2: $\hat{\gamma}_0 = \gamma_m(m/(m \quad + 1))$

Calculating $\boldsymbol{\gamma}_m$ for different m and $\boldsymbol{\gamma}_0$ combinations using Equation (14), one will find the following approximation stands well for $m \geq 10$:

$$\gamma_m \approx \frac{m+1}{m}\gamma_0 . \quad (17)$$

From Equation (17), we obtain the following method of estimating $\boldsymbol{\gamma}_0$ from $\boldsymbol{\gamma}_m$:

$$\hat{\gamma}_0 = \frac{m}{m+1}\gamma_m . \quad (18)$$

For those who may be interested, this method may be proven by resolving $\boldsymbol{\gamma}_0$ from Equation (14) using Taylor series expansion approximation. An advantage of this method is that one could use it to have a more accurate FMI estimation from the m and the $\boldsymbol{\gamma}_m$ information available in an earlier publication that uses a small m and $\boldsymbol{\gamma}_m$ to estimate FMI.

4.3. Improved method 3: $\hat{\gamma} = c_m/(c_m + 1)$, where $c_m = B_m/U_m$

In Section 2.2, we proved that $E(B_m/U_m) = B_0/U_0$. In other words, B_m/U_m is an unbiased estimation of B_0/U_0. As a result, Equation (19) below is a better $\boldsymbol{\gamma}_0$ estimation than $\hat{\gamma}_0 = \gamma_m$:

$$\hat{\gamma}_0 = \frac{c_m}{1+c_m} . \quad (19)$$

where $c_m = B_m/U_m$. Harel used this method to estimate γ_0 for his study on two-stage MI (Harel, 2007). However, the justification for this method discussed here was not available in Harel's paper or any other published literature (Harel, 2007).

5. Results from MI trials of PWS12

5.1. Methods

PWS was a supplemental survey of NAMCS, which collects data about the provision and use of ambulatory medical care services in the United States (Lau, McCaig, & Hing, 2016). The 2012 PWS data (PWS12) were used for the MI trial, which had 2,567 responded physicians in the sample. PWS data can be accessed via NCHS Research Data Center (RDS) program (https://www.cdc.gov/rdc/index.htm).

MI was conducted on three variables representing the physician's practice size at different scales, namely SIZE100, SIZE20, and SIZE5. The three variables had the same missing data percentage of 29% due to item non-responses. The hot-deck imputation method (Andridge & Little, 2010) was used. The RDS-released PWS12 data, which had 3.6% of missing values for SIZE after some of the missing values in PWS12 were replaced by the corresponding non-missing values for the same physician from the 2011 PWS data, were used as the hot-deck donor. Two MI models denoted as MI-1 and MI-2 were used. MI-1 did not use any covariate in the imputation and the non-missing replacement values for the missing value were randomly chosen from entire donor dataset. MI-2 used PRIMEMM as the covariate in the imputation and the non-missing replacement values for the missing value were randomly chosen from the cell of the same PRIMEMM value in donor dataset. PRIMEMM was the physician's primary employment type that was coded into nine categories for this research. The MIs had $m = 3$, 5, 10, 20, 30, 40, 60, 80, and 100, with each MI being repeated for 30 times. Excluding m, there were 12 treatment combinations (3 imputed variables × 2 imputation models × 2 analytic models). The hot-deck imputation method used in this study was similar to that used by the survey for creating the RDC-released PWS12 data. According to Rubin (1987, equation 4.3.8), the hot-deck bias can be expressed as $E(B) = B(n1/n)$, where n is the number of the units of the full sample and $n1$ is the number of the units with observed values. Since the $n1/n$ ratio is independent of m, the percentage fraction of the hot-deck bias would be a fixed value as long as the $n1/n$ ratio is fixed. Therefore the m–γ_m relationship obtained from the hot-deck-based MI trials should still be valid. One should be aware of the potential hot-deck bias when interpreting the results of this study.

The quantity of interest (Q) was the means of the SIZE100, SIZE20, and SIZE5. Two analytical models denoted as Anal-1 and Anal-2 were used. In Anal-1, U_i, the within-imputation variance of the ith complete dataset generated by the MI, was the total variance of SIZE100, SIZE20, or SIZE5 in the ith dataset. In Anal-2, U_i was the variance of the ith dataset after the variance due to the effect of PRIMEMM was removed. Analyses were based on un-weighted data. Results obtained in this study were for research purpose only.

Barnard and Rubin (1999) suggested that, for making the statistical inferences in MI-involved analyses, instead of using the degrees of freedom (v) as defined by Equation (6), the adjusted degrees of freedom (DFa) as proposed by their paper should be used where the

complete-data degrees of freedom is not sufficiently large. However in the γ_m definition, i.e. Equation (7), v does not function as the degrees of freedom per se but mealy as a mathematical value in the estimation of γ_m. We have found that replacing v in Equation (7) with DFa will result in an erroneous estimation of γ_m. Therefore we used v, instead of DFa, when used Equation (7) for the γ_m estimation in this study.

5.2. The γ_m decrease with the increase of m in the MI trials

Would the γ_m decrease due to the increase of m (see Section 2.1 and 3.1) be big enough to stand out from sampling errors and other noises in real-world MI analyses? The answer is yes, as demonstrated by the data in Figure 3. Figure 3 shows the effects of m on γ_m in SIZE100, SIZE20, and SIZE5 for the two MI models for Anal-2. In spite of the γ_m variations due to sampling errors as shown by the error bars in the graphs, the dominant trend was clear: γ_m decreased significantly as m increased from 3 to 100. The γ_m values at m = 3–40 were significantly greater than γ_{100} in most cases (Figure 3). These results suggest that the γ_m decrease with the increase of m is not ignorable in FMI estimation in real world data analyses.

5.3. Variation of γ_m, B_m, and U_m

In establishing the MI framework, Rubin (1987) assumed that $U_m \approx U_0$, which would be more likely to be true if the variance of U_m is negligible. The authors did not find any information on the magnitude of U_m variance in published literature. A detailed study on B_m variance was reported by Pan et al. (Pan et al., 2014). The variance of B_m was substantial when $m < 30$ (Pan et al., 2014). The variations in B_m and U_m would inevitably lead to γ_m variation. As a result, when using $\hat{\gamma}_0 = \gamma_m$ at an insufficient m, the inaccuracy of $\hat{\gamma}_0$ would not only come from $E(\gamma_m) > \gamma_0$ but also from the variation of γ_m The possible bias from sampling-error-driven γ_m variation has not been given an adequate attention.

The coefficient of variations (CV) of B_m, U_m, and γ_m are presented in Table 2. Both the imputations models and the analytic models affected the variations of γ_m, B_m, and U_m (Table 2). CV of U_m was much smaller—usually 1–10% that of B_m. The CV of B_m and γ_m were very similar, with the CV of γ_m being always slightly smaller than that of B_m. The greater the m, the smaller the variations of γ_m, B_m, and U_m (Table 2). These results were in agreement with Harel's conclusion (Hogg et al., 2013) that it is necessary to choose a sufficient m for MI to control the variations of γ_m. Due to the significant effects of the MI model and the analytic model on the variations of γ_m, B_m, and U_m (Table 2), it may not be possible to propose a single m that fits all situations for controlling the variance of γ_m, B_m, and U_m.

An advantage of using $\hat{\gamma}_0 = \gamma_{m \geq 100}$ is that not only can this method reduce the $E(\gamma_m) > \gamma_0$ bias but also reduce γ_m variation because of a large m. The other two improved methods can effectively reduce or even eliminate the $E(\gamma_m) > \gamma_0$ bias even if when m is small. However the $\hat{\gamma}_0$ inaccuracy may be a concern for any FMI estimation methods unless a sufficient m is chosen. Data in Figure 3 suggest that a ≥20 m may be necessary to reduce the γ_m variation to an acceptable level for using $\hat{\gamma}_0 = \gamma_m(m/(m+1))$ and $\hat{\gamma}_0 = c_m/(c_m+1)$.

5.4. Comparison of different FMI estimation methods

Table 3 presents data for visualizing the performance of these three improved γ_0 estimation methods described in Section 4 in comparison with the default method $\hat{\gamma}_0 = \gamma_m$ in an example of real-world data analyses. The treatment combination of the MI trials was {SIZE20, MI-2, Anal-2}. The control values was the γ_m values at $m = 3, 5$, etc., which would be the FMI estimation when the default method was used. The γ_{100} value was used as the $\hat{\gamma}_0$ for the improved method $\hat{\gamma}_0 = \gamma_{m \geq 100}$. The best $\hat{\gamma}_0$ was calculated by Equation (19) using $(\bar{B}_{100})/(\bar{U}_{100})$ as the estimate of B_0/U_0, where $\bar{B}_{100}$ and $\bar{U}_{100}$ were the mean of the 30 replicates of B_{100} and U_{100}.

For $m \leq 80$, all three improved methods performed better than the control method (Table 3). These results suggest that the three improved methods proposed in this paper can be used to replace the control method in real world data analyses. In general we recommend to use $\hat{\gamma}_0 = c_m/(1 + c_m)$, for it essentially eliminates the $E(\gamma_m) > \gamma_0$ bias at all levels of m. But the two other methods may come in handy under certain circumstances. For example, if an earlier publication which had used $m = 5$ without providing B_m and U_m values, one can simply use $\hat{\gamma}_0 = \gamma_m m/(1 + m)$ to convert the biased γ_0 estimate of the paper into a more correct γ_0 estimate.

6. Conclusions

In most published researches, γ_∞ and γ_0 are treated as synonyms. However, the two are different. The γ_0 is independent of MI, whereas γ_∞ is a parameter of MI. γ_∞ equals to γ_0 only if $B_m/U_m = B_0/U_0$. To use MI for FMI estimation, one has to assume $\gamma_\infty = \gamma_0$, which will also be the assumption here.

The γ_m decreases with the increase of m. We quantified the m–γ_m relationship. The magnitude and the rate of the γ_m decrease varies with m and γ_0. At $m = 2$, γ_2 is greater than γ_0 by 50–81% depending on the γ_0 level. At $m = 5$, the recommended m value as being sufficient by some (e.g. Rubin, 1987), γ_m is greater than γ_0 by 20–25% when γ_0 value ranges from 0.001 to 0.6. The decrease of γ_m with the increase of m determines that $E(\gamma_m) > \gamma_0$ for any finite m. The results from the MI trials suggest that the volume of the γ_m decrease with increased m is not ignorable in real world data analyses in spite of the noises from sampling errors and other sources.

$E(B_m)$ and $E(U_m)$ are independent of m. Therefore, the decrease of γ_m with the increase of m is not due to an indirect effect of m on $E(B_m)$ and $E(U_m)$. As a result, it is not necessary to use the B_m and U_m from the same MI for best γ_m estimation. Instead, one should use the best estimates of B_0 and U_0 available, which leads to the development of Equation (14) that links γ_m to γ_0 directly.

The variation in γ_m can be substantial. The CV of γ_m was essentially identical with that of B_m, and CV of U_m was 1–10% that of γ_m or B_m. The variation of γ_m is smaller as m gets bigger. The inaccuracy of FMI estimation due to γ_m variation should be concerned in FMI estimation when m is small regardless what method is used.

The current method $\hat{\gamma}_0 = \gamma_m$ may result in a substantial FMI overestimation when m is not sufficiently large. Three improved methods are proposed for estimating γ_0 from MI of a finite m. These three methods are (1)$\hat{\gamma}_0 = \gamma_{m \geq 100}$, (2)$\hat{\gamma}_0 = \gamma_m(m/(m+1))$, and (3) $\hat{\gamma}_0 = c_m/(c_m + 1)$, where $c_m = B_m/U_m$. In our MI trials, all three improved methods gave more accurate γ_0 estimates than $\hat{\gamma}_0 = \gamma_m$ where m is less than 80.

When m is sufficiently large, say, $m \geq 100$, all three methods should give a statistically sound estimation of γ_0. When m is not sufficiently large, say, $m < 100$, the third method $\hat{\gamma}_0 = c_m/(c_m + 1)$ should be one's best option for γ_0 estimation. The second method $\hat{\gamma}_0 = \gamma_m(m/(m+1))$ has its value where B_m and U_m are not available and the only values available to use for γ_0 estimation are m and γ_m.

Acknowledgements

Dr. Yulei He of NCHS is sincerely thanked for his invaluable insights and suggestions.

Funding

This work was performed while the authors were under employment by the US federal government. The authors used government resources, e.g. computers and office spaces, for this research. The authors did not receive any specified funding for this research from inside or outside US government resources.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the National Center for Health Statistics (NCHS) or Centers for Disease Control (CDC), USA.

References

Andridge RR, & Little RJ (2010). A review of hot deck imputation for survey non-response. International Statistical Review, 78(1), 40–64. doi:10.1111/j.1751-5823.2010.00103.x [PubMed: 21743766]

Barnard J, & Rubin DB (1999). Small-sample degrees of freedom with multiple imputation. Biometrika, 86(4), 948–955. doi:10.1093/biomet/86.4.948

Bodner TE (2008). What improves with increased missing data imputations? Structural Equation Modeling: A Multidisciplinary Journal, 15, 651–675. doi:10.1080/10705510802339072

Carpenter J, & Kenward M (2013). Multiple imputation and its application John Wiley & Sons.

Dohoo IR (2015). Dealing with deficient and missing data. Preventive Veterinary Medicine, 122(1–2), 221–228. doi:10.1016/j.prevetmed.2015.04.006 [PubMed: 25930986]

Graham JW, Olchowski AE, & Gilreath TD (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science, 8(3), 206–213. doi:10.1007/s11121-007-0070-9 [PubMed: 17549635]

Harel O (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. Statistical Methodology, 4, 75–89. doi:10.1016/j.stamet.2006.03.002

He Y, Shimizu I, Schappert S, Xu J, Beresovsky V, Khan D, … Schenker N. (2016). A note on the effect of data clustering on the multiple-imputation variance estimator: A theoretical addendum to the Lewis et al. article in JOS 2014. Journal of Official Statistics, 32(1), 147–164. doi:10.1515/jos-2016-0007

Hershberger SL, & Fisher DG (2003). A note on determining the number of imputations for missing data. Structural Equation Modeling, 10(4), 648–650. doi:10.1207/S15328007SEM1004_9

Hogg RV, McKean JW, & Craig TA (2013). Introduction to mathematical statistics (7th ed.). Boston: Pearson Education, Inc.

Khare M, Little RJA, Rubin DB, & Schafer JL 1993 Multiple imputation of NHANES III. Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 297–302.

Lau DT, McCaig LF, & Hing E (2016). Toward a more complete picture of outpatient, office-based health care in the U.S: Expansion of NAMCS. American Journal of Preventive Medicine, 51(3), 403–409. doi:10.1016/j.amepre.2016.02.028 [PubMed: 27079637]

Lewis T, Goldberg E, Schenker N, Beresovsky V, Schappert S, Decker S, … Shimizu I (2014). The relative impacts of design effects and multiple imputation on variance estimates: A case study with the 2008 National Ambulatory Medical Care Survey. Journal of Official Statistics, 30(1), 147–161. doi:10.2478/jos-2014-0008

Little RJ, Wang J, Sun X, Tian H, Suh E-Y, Lee M, … Mohanty S (2016). The treatment of missing data in a large cardiovascular clinical outcomes study. Clinical Trials, 13(3), 344–351. doi: 10.1177/1740774515626411 [PubMed: 26908543]

Pan Q, Wei R, Shimizu I, & Jamoom E (2014). Determining sufficient number of imputations using variance of imputation variances: Data from 2012 NAMCS Physician Workflow Mail Survey. Applied Mathematics, 5, 3421–3430. doi:10.4236/am.2014.521319 [PubMed: 27398258]

Rezvan PH, Lee KJ, & Simpson JA (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. doi:10.1186/s12874-015-0022-1 [PubMed: 25880850]

Royston P (2004). Multiple imputation of missing values. The Stata Journal, 4(3), 227–241.

Rubin DB (1987). Multiple imputation for nonresponse in surveys New York, NY: John Wiley & Sons.

Savalei V, & Rhemtulla M (2012). Missing information from full information maximum likelihood. Structural Equation Modeling: A Multidisciplinary Journal, 19(3), 477–494. doi: 10.1080/10705511.2012.687669

Schafer JL (2001). Analyzing the NHANES III multiply imputed data set: Methods and examples Retrieved from ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHANES/NHANESIII/7a/doc/analyzing.pdf.

Schenker N, Raghunathan TE, Chiu P-L, Makuc DM, Zhang G, & Cohen AJ (2006). Multiple imputation of missing income data in the national health interview survey. Journal of the American Statistical Association, 101, 924–933. doi:10.1198/016214505000001375

Serfling R (1980). Approximation theorems of mathematical statistics John Wiley & Sons, Inc.

Siddique J, Harel O, Crespic CM, & Hedekerd D (2014). Binary variable multiple-model multiple imputation to address missing data mechanism uncertainty: Application to a smoking cessation trial. Statistics in Medicine, 33, 3013–3028. doi:10.1002/sim.6137 [PubMed: 24634315]

Van Buuren S (2012). Flexible imputation of missing data Boca Raton, FL: Chapman and Hall/CRC Press.

Wagner J (2010). The fraction of missing information as a tool for monitoring the quality of survey data. Public Opinion Quarterly, 74(2), 223–243. doi:10.1093/poq/nfq007

Zheng T, & Lo S-H (2008). Comment: Quantifying the fraction of missing information for hypothesis testing in statistical and genetic studies. Statistical Science, 23, 321–324. doi:10.1214/08-STS244A

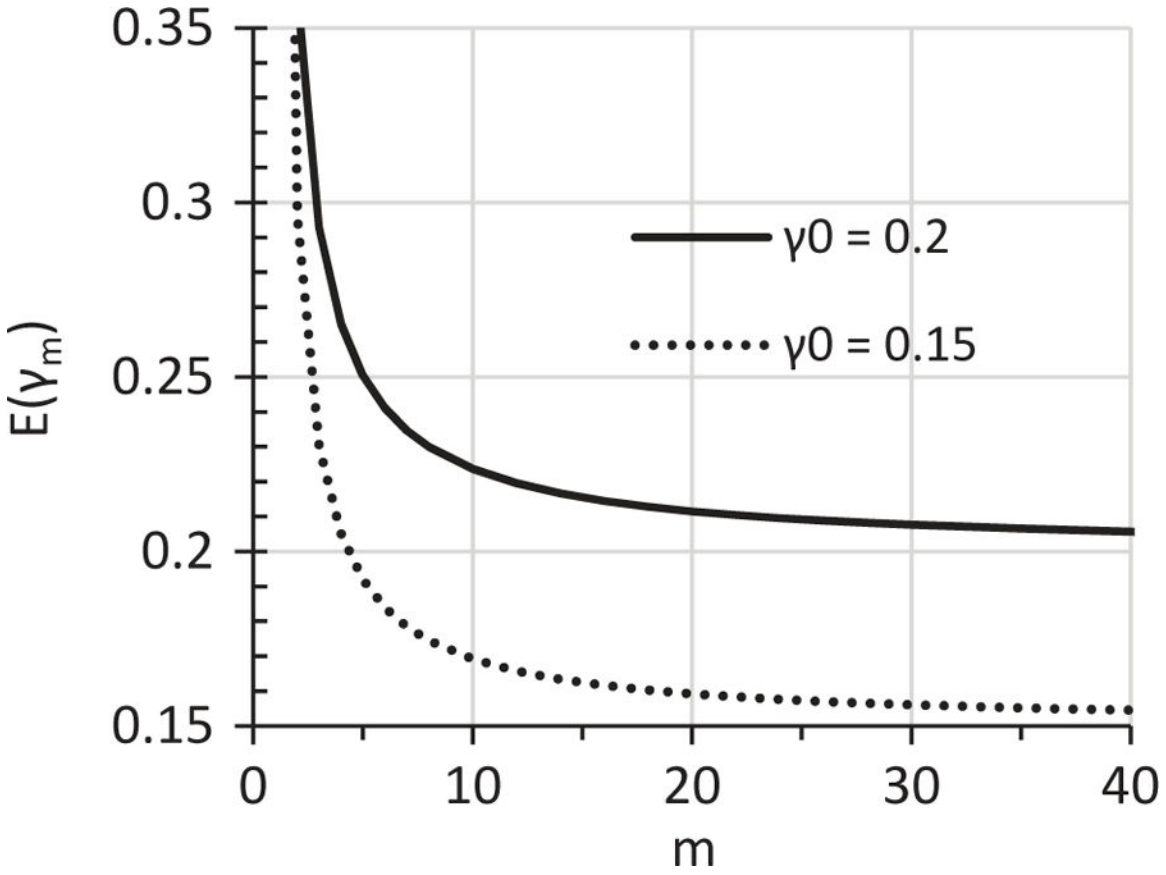

Figure 1.
The m–$E(\gamma_m)$ relationship curve at $\gamma_0 = 0.2$ and 0.15 as determined by Equation (14).

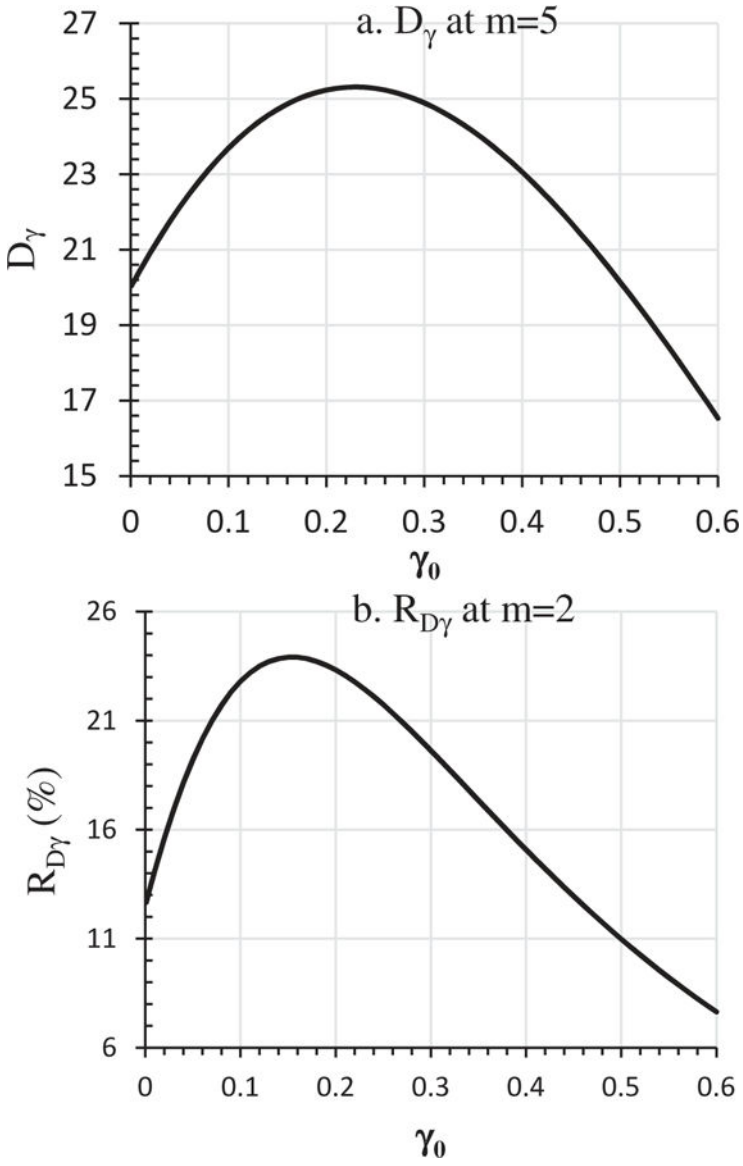

Figure 2.

Effects of γ_0 levels on D_γ as defined by Equation (15) and $R_{D\gamma}$ as defined by Equation (16): a. D_γ at $m = 5$; b. D_γ at $m = 2$.

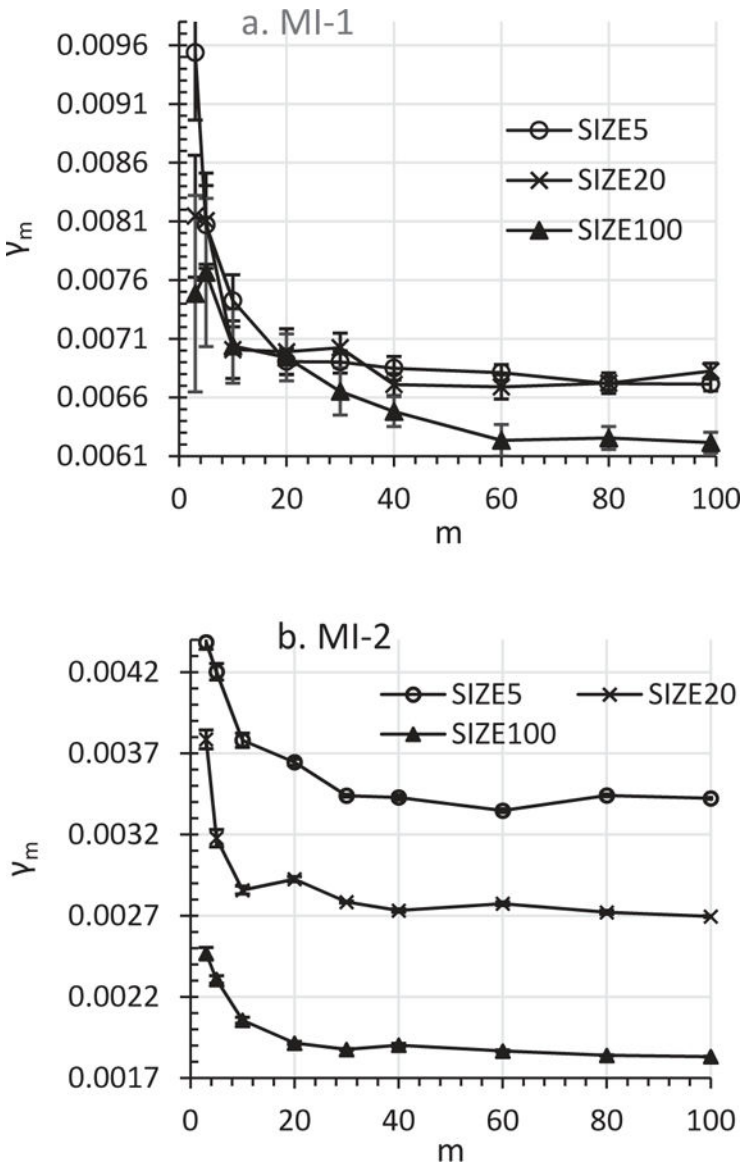

Figure 3.

Effects of m on γ_m at δ = 29% for analytic model = Anal-2: MI model = MI-1; b. MI model = MI-2.

Table 1.

Changes of $E(\gamma_m)$, D_γ, and $R_{D\gamma}$ with the increase of m at different γ_0 levels, where $D_\gamma = 100(E(\gamma_m)-\gamma_0)/\gamma_0$ and $R_{D\gamma} = 100(\gamma_m - \gamma_{m+1})/\gamma_{m+1}$

m	$E(\gamma_m)$		$R_{D\gamma}$		D_γ	
	$\gamma_0 = 0.2$	$\gamma_0 = 0.01$	$\gamma_0 = 0.2$	$\gamma_0 = 0.01$	$\gamma_0 = 0.2$	$\gamma_0 = 0.01$
2	0.361	0.01536	23.33	14.12	80.59	53.64
5	0.250	0.01205	3.872	2.957	25.23	20.47
10	0.224	0.01102	1.0240	0.8502	11.84	10.16
20	0.212	0.01051	0.2664	0.2306	5.750	5.062
40	0.206	0.01025	0.0682	0.0602	2.837	2.528
60	0.204	0.01017	0.0306	0.0272	1.883	1.684
100	0.202	0.01010	0.0111	0.0099	1.126	1.010
200	0.201	0.01005	0.0028	0.0025	0.561	0.505

Table 2.

Coefficient of variations (%) of B_m, U_m, and γ_m for SIZE100

m	MI-1, Anal-1			MI-2, Anal-2		
	B_m	U_m	γ_m	B_m	U_m	γ_m
3	20.24	0.185	20.19	1.52	0.0553	1.50
5	11.12	0.179	11.10	1.00	0.0200	1.01
10	7.75	0.135	7.75	0.88	0.0353	0.91
20	5.91	0.098	5.89	0.49	0.0282	0.48
30	6.24	0.077	6.26	0.24	0.0150	0.24
40	3.60	0.048	3.59	0.61	0.0180	0.59
60	2.74	0.058	2.73	0.39	0.0091	0.39
80	2.48	0.041	2.47	0.17	0.0062	0.17
100	2.45	0.037	2.45	0.15	0.0081	0.16

Table 3.

Comparison of different γ_0 estimation methods for SIZE20 with imputation model = MI-2 and analytic model = Anal-2 in the PWS12 MI trials. The best $\hat{\gamma}_0$ was calculated by Equation (19) using $(\bar{B}_{100})/(\bar{U}_{100})$ as the estimate of B_0/U_0, where $\bar{B}_{100}$ and $\bar{U}_{100}$ were the mean of the 30 replicates of B_{100} and U_{100}, respectively

m	SIZE20, MI-2, Anal-2			
	Control	Improved		
	$\hat{\gamma}_0 = \gamma_m$	$\hat{\gamma}_0 = \gamma_{m \geq 100}$	$\hat{\gamma}_0 = c_m/(1+c_m)$	$\hat{\gamma}_0 = \gamma_m(m/(1+m))$
3	0.00379		0.00283	0.00284
5	0.00318		0.00265	0.00265
10	0.00286		0.00260	0.00260
20	0.00293		0.00279	0.00279
30	0.00278		0.00269	0.00269
40	0.00273		0.00267	0.00267
60	0.00277		0.00273	0.00273
80	0.00272		0.00269	0.00269
100	0.00270	0.00270	0.00267	0.00267
(∞)	Best $\hat{\gamma}_0$: 0.00267			