Improved methods for estimating fraction of missing information in multiple imputation

Qiyuan Pan1* and Rong Wei2

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 \( \gamma_m \), the sample value of FMI estimated from MI of a limited \( m \), is used as the estimate of \( \gamma_0 \), 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 \( \gamma_m \) would inevitably overestimate \( \gamma_0 \). 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

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 impact of missing data.

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.
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 \( \gamma_0 \) 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 Hedeker (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 \( \gamma_0 \) and \( \gamma_m \) is the FMI obtained from MI of a given \( m \), e.g. (Khare, Little, Rubin, & Schafer, 1993; Lewis et al., 2014; 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:

\[
Q_m = \frac{1}{m} \sum_{i=1}^{m} Q_i, \quad (1)
\]

where subscript \( m \) stands for a finite and infinite \( m \), the subscript \( 0 \) for the population value, the subscript \( i \) for the \( i \)th imputation, and the bar hat for the parameter’s mean.

\[
B_m = \frac{1}{m-1} \sum_{i=1}^{m} (Q_i - Q_m)^2 \quad (2)
\]

\[
U_m = \frac{1}{m} \sum_{i=1}^{m} U_i \quad (3)
\]

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

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

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

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 $\gamma_0$ in practice because $B_0$, $U_0$, and $T_0$ are usually unknown. No researchers have provided an equation that explicitly links $\gamma_m$ and $\gamma_0$. The justification for using $\gamma_0 = \gamma_m$ is not available from Equations (1)–(11).

Assume $\gamma_m = \gamma_0$. 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 $\gamma_m$ from each MI repeat as $\gamma_{m1}$, $\gamma_{m2}$, $\gamma_{m3}$, ..., $\gamma_{mj}$. 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:

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

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

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

For FMI estimation, Harel’s 2007 paper (Harel, 2007) is important in that it pointed out that, unlike $\gamma_m$ 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 $\gamma_0 = B_m/(U_m + B_m)$ 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 $\gamma_0 = B_m/(U_m + B_m)$ 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 $m$, $\gamma_m$, $\gamma_m$, and $\gamma_0$, quantified the decrease of $\gamma_m$, 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 multivariate 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 $m$, $\gamma_m$, $\gamma_m$, and $\gamma_0$

2.1. The condition for $\gamma_m = \gamma_0$

To use $\gamma_m$ for estimating $\gamma_0$, one must assume $\gamma_m = \gamma_0$. Most researchers simply treat $\gamma_m$ and $\gamma_0$ 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 $\gamma_m$ and $\gamma_m$, respectively. The population for $\gamma_0$, however, is not POI but the population of the sampling units of the survey. A $\gamma_m$ is inseparably linked to an MI, but a $\gamma_0$ can be independent of MI. The $\gamma_0$ 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 $\gamma_m = \gamma_0$ is $B_m/U_m = B_0/U_0$. When we use MI to estimate $\gamma_0$, we
have to assume \( y_m = y_0 \), which is probably why \( y_m \) and \( y_0 \) are often treated as synonyms in MI analyses. In this paper, we will assume \( y_m = y_0 \) because we use MI to estimate \( y_0 \).

2.2. \( B_m/U_m \) is independent of \( m \)

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

\[
y_m = \frac{(1 + \frac{1}{m}) \frac{B_m}{U_m} + 2 \left( (m - 1) \left( 1 + \frac{1}{(1 + \frac{1}{m})^2} \right)^2 + 3 \right)}{\left( 1 + \frac{1}{m} \right) \frac{B_m}{U_m} + 1}.
\] (13)

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

Whether \( c_m \) is independent of \( m \) is important in understanding the \( m \)-\( y_m \) relationship. If \( c_m \) depends on \( m \), the direct effects of \( m \) on \( y_m \) would be confounded by the indirect effects of \( m \) on \( y_m \) 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 \( m \)-\( y_m \) 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 \( x \) 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 \( m \)-\( y_m \) 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 \( y_m \) become a function of the single factor \( m \), i.e. \( y_m = F(m) \).

2.3. The \( y_m = F(m,y_0) \) equation

When \( m \) goes infinite, \( y_m \) becomes \( y_0 \). Our goal is to establish the mathematic relationship between \( y_m \) and \( y_0 \) 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 \( m \)-\( y_m \) relationship because \( c_m \) is independent of \( m \). What is the best value to choose for \( c \) to obtain the most truthful \( m \)-\( y_m \) 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 \( m \)-\( y_m \) relationship as determined by Equation (13) would reflect the true \( m \)-\( y_m \) relationship. From Equations (10) and (11) we can obtain \( B_0/U_0 = y_0/(1 – y_0) \). Replacing \( B_m/U_m \) in Equation (13) with \( y_0/(1 – y_0) \), we obtain an equation that directly links \( y_m \) to \( y_0 \) as follows:

\[
y_m = E(y_m) = F(m, y_0) = \frac{(1 + \frac{1}{m}) \frac{y_0}{1-y_0} + 2 \left( (m - 1) \left( 1 + \frac{1}{(1 + \frac{1}{m})^2} \right)^2 + 3 \right)}{\left( 1 + \frac{1}{m} \right) \frac{y_0}{1-y_0} + 1}.
\] (14)

Establishment of equation is a significant step forward in understanding the relationship between \( m \), \( y_m \), and \( y_0 \) because it links the three factors in the same equation for any \( m \), finite or infinite.

For a given analysis of a given dataset, \( y_0 \) is a constant. When we repeat the same MI of a given \( m \) for \( j \) times, the \( y_m \) value from each repeat of the MI will not change when the \( y_m \) is determined by Equation (14). In other words, we will have \( y_{m1} = y_{m2} = y_{m3} = \ldots y_{mj} = E(y_m) \) (see Equation (12)). In other words, the \( y_m \) value obtained from Equation (14) will be \( E(y_m) \). For different data and
analyses, γ₀ is a variable. Equation (14) shows that \( E(\gamma_m) \) is a function of the two factors, \( m \) and \( \gamma_0 \), 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 \( \mu \), which provides the theoretical base for \( \mu = \bar{x} \) (Hogg et al., 2013). The use of \( \gamma_0 = \gamma_m \) implies the assumption that \( E(\gamma_m) = \gamma_0 \). Using Equation (14), the \( m-\gamma_m \) relationship curve can be constructed for any given \( \gamma_0 \).

Figure 1 presents the \( m-\gamma_m \) relationship curve at \( \gamma_0 = 0.2 \) and 0.15 as determined by Equation (14).

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 \( \gamma_0 = \gamma_m \) must be biased. One achievement of this paper is that we successfully quantified the bias of the current FMI estimation method. We use \( D_\gamma \), the percentage difference between \( E(\gamma_m) \) and \( \gamma_0 \) as the parameter to measure this bias, i.e.:

\[
D_\gamma = 100 \frac{\gamma_m - \gamma_0}{\gamma_0}.
\] (15)

Table 1 presents the \( D_\gamma \) values at different \( \gamma_0 \) and \( m \) values as determined by Equation (14). At a given \( m \), \( D_\gamma \) differs at different \( \gamma_0 \) values (Table 1). For \( m = 2 \), \( D_\gamma \) is 80.59% and 53.64% for \( \gamma_0 = 0.2 \) and 0.01, respectively (Table 1). When \( \gamma_0 \) increased from 0.001 to 0.6, \( D_\gamma \) first increases with the increase of \( \gamma_0 \), reaches a peak, and then decreases (Table 1 and Figure 2). The value of the \( \gamma_0 \) at which \( D_\gamma \) reaches the peak differs with \( m \) (data not shown). For \( m = 5 \), the maximum \( D_\gamma \) value of 25.31% occurs at \( \gamma_0 = 0.23 \), and the minimum \( D_\gamma \) 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 $\gamma_m$ decrease rate: smaller at larger $m$

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

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

<table>
<thead>
<tr>
<th>$m$</th>
<th>$E(\gamma_m)$</th>
<th>$R_{\gamma}$</th>
<th>$D_{\gamma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$\gamma_0 = 0.2$</td>
<td>0.361</td>
<td>23.33</td>
</tr>
<tr>
<td></td>
<td>$\gamma_0 = 0.01$</td>
<td>0.01536</td>
<td>14.12</td>
</tr>
<tr>
<td>5</td>
<td>$\gamma_0 = 0.2$</td>
<td>0.250</td>
<td>3.872</td>
</tr>
<tr>
<td></td>
<td>$\gamma_0 = 0.01$</td>
<td>0.01205</td>
<td>2.957</td>
</tr>
<tr>
<td>10</td>
<td>$\gamma_0 = 0.2$</td>
<td>0.224</td>
<td>1.0240</td>
</tr>
<tr>
<td></td>
<td>$\gamma_0 = 0.01$</td>
<td>0.01102</td>
<td>0.8502</td>
</tr>
<tr>
<td>20</td>
<td>$\gamma_0 = 0.2$</td>
<td>0.212</td>
<td>0.2664</td>
</tr>
<tr>
<td></td>
<td>$\gamma_0 = 0.01$</td>
<td>0.01051</td>
<td>0.2306</td>
</tr>
<tr>
<td>40</td>
<td>$\gamma_0 = 0.2$</td>
<td>0.206</td>
<td>0.0682</td>
</tr>
<tr>
<td></td>
<td>$\gamma_0 = 0.01$</td>
<td>0.01025</td>
<td>0.0602</td>
</tr>
<tr>
<td>60</td>
<td>$\gamma_0 = 0.2$</td>
<td>0.204</td>
<td>0.0306</td>
</tr>
<tr>
<td></td>
<td>$\gamma_0 = 0.01$</td>
<td>0.01017</td>
<td>0.0272</td>
</tr>
<tr>
<td>100</td>
<td>$\gamma_0 = 0.2$</td>
<td>0.202</td>
<td>0.0111</td>
</tr>
<tr>
<td></td>
<td>$\gamma_0 = 0.01$</td>
<td>0.01010</td>
<td>0.0099</td>
</tr>
<tr>
<td>200</td>
<td>$\gamma_0 = 0.2$</td>
<td>0.201</td>
<td>0.0028</td>
</tr>
<tr>
<td></td>
<td>$\gamma_0 = 0.01$</td>
<td>0.01005</td>
<td>0.0025</td>
</tr>
</tbody>
</table>

Figure 2. Effects of $\gamma_0$ levels on $D_{\gamma}$ as defined by Equation (15) and $R_{\gamma}$ as defined by Equation (16): a. $D_{\gamma}$ at $m = 5$; b. $D_{\gamma}$ at $m = 2$.
\( R_{Dy} \) is affected by both \( m \) and \( \gamma_0 \) (Table 1 and Figure 2, b1 and b2). At \( m = 5 \), \( R_{Dy} \) is 3.87% and 2.96% for \( \gamma_0 = 0.2 \) and 0.01, respectively (Table 1). Figure 2(b) show that \( R_{Dy} \) increases initially, reaches a peak, and then decreases as \( \gamma_0 \) increases from 0.001 to 0.6. For \( m = 2 \), the maximum \( R_{Dy} \) = 23.91% occurs at \( \gamma_0 = 0.15 \) (Figure 2(b)). For \( m = 5 \), the maximum \( R_{Dy} \) = 3.88% occurs at \( \gamma_0 = 0.21 \) (data not shown). The gradual reduction of \( R_{Dy} \) makes it possible for choosing a sufficient \( m \) when the \( m \)-driven \( \gamma_m \) reduction becomes negligibly small.

4. Improved methods for \( \gamma_0 \) 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 \( E(\gamma_m) \) approaches \( \gamma_0 \) as \( m \) gets larger. Therefore, \( \gamma_m \) would estimate \( \gamma_0 \) 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 \( \gamma_0 \) using \( \hat{\gamma}_0 = \gamma_m \) offers another criterion for determining a sufficient \( m \). As measured by \( R_{Dy} \), 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 \( D_0 \) would be about 1% or less for any reasonable \( \gamma_0 \) values when \( m \) is greater than 100. We arbitrarily choose a bias of \( \leq 1\% \) as an acceptable level and recommend \( m \geq 100 \) as being sufficient for an adequately accurate estimation of \( \gamma_0 \) using \( \hat{\gamma}_0 = \gamma_m \). This method can be expressed as \( \hat{\gamma}_0 = \gamma_{m \geq 100} \).

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

Calculating \( \gamma_m \) for different \( m \) and \( \gamma_0 \) combinations using Equation (14), one will find the following approximation stands well for \( m \geq 10 \):

\[
\gamma_m \approx \frac{m + 1}{m} \gamma_0. \tag{17}
\]

From Equation (17), we obtain the following method of estimating \( \gamma_0 \) from \( \gamma_m \):

\[
\hat{\gamma}_0 = \frac{m}{m + 1} \gamma_m. \tag{18}
\]

For those who may be interested, this method may be proven by resolving \( \gamma_0 \) 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 \( \gamma_m \) information available in an earlier publication that uses a small \( m \) and \( \gamma_m \) to estimate FMI.

4.3. Improved method 3: \( \hat{\gamma}_0 = c_m/(c_m + 1) \), 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 \( \gamma_0 \) estimation than \( \hat{\gamma}_0 = \gamma_m \):

\[
\hat{\gamma}_0 = \frac{c_m}{1 + c_m}. \tag{19}
\]

where \( c_m = B_m/U_m \). Harel used this method to estimate \( \gamma_0 \) 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 \( \times 2 \) imputation models \( \times 2 \) 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 \( m=\gamma_m \) 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 \( i \)th complete dataset generated by the MI, was the total variance of SIZE100, SIZE20, or SIZE5 in the \( i \)th dataset. In Anal-2, \( U_i \) was the variance of the \( i \)th 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 \( \gamma_m \) 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 \( \gamma_m \). We have found that replacing \( v \) in Equation (7) with DFa will result in an erroneous estimation of \( \gamma_m \). Therefore we used \( v \), instead of DFa, when used Equation (7) for the \( \gamma_m \) estimation in this study.

5.2. The \( \gamma_m \) decrease with the increase of \( m \) in the MI trials

Would the \( \gamma_m \) 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 \( \gamma_m \) in SIZE100, SIZE20, and SIZE5 for the two MI models for Anal-2. In spite of the \( \gamma_m \) variations due to sampling errors as shown by the error bars in the graphs, the dominant trend was clear: \( \gamma_m \) decreased significantly as \( m \) increased from 3 to 100. The \( \gamma_m \) values at \( m = 3 – 40 \) were significantly greater than \( \gamma_{100} \) in most cases (Figure 3). These results suggest that the \( \gamma_m \) decrease with the increase of \( m \) is not ignorable in FMI estimation in real world data analyses.

5.3. Variation of \( \gamma_m \), \( B_m \), and \( U_m \)

In establishing the MI framework, Rubin (1987) assumed that \( U_m = U_0 \) 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 \( \gamma_m \) variation. As a result, when using \( \gamma_0 = \gamma_m \) at an insufficient
m, the inaccuracy of \( \gamma_0 \) would not only come from \( E(\gamma_m) > \gamma_0 \) but also from the variation of \( \gamma_m \). The possible bias from sampling-error-driven \( \gamma_m \) variation has not been given an adequate attention.

The coefficient of variations (CV) of \( B_m, U_m, \) and \( \gamma_m \) are presented in Table 2. Both the imputations models and the analytic models affected the variations of \( \gamma_m, B_m, \) 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 \( \gamma_m \) were very similar, with the CV of \( \gamma_m \) being always slightly smaller than that of \( B_m \). The greater the \( m \), the smaller the variations of \( \gamma_m, B_m, \) 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 \( \gamma_m \). Due to the significant effects of the MI model and the analytic model on the variations of \( \gamma_m, B_m, \) and \( U_m \) (Table 2), it may not be possible to propose a single \( m \) that fits all situations for controlling the variance of \( \gamma_m, B_m, \) and \( U_m \).

An advantage of using \( \hat{\gamma}_0 = \gamma_m(m/(m + 1)) \) is that not only can this method reduce the \( E(\gamma_m) > \gamma_0 \) bias but also reduce \( \gamma_m \) 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 \( \geq 20 \) \( m \) may be necessary to reduce the \( \gamma_m \) 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 \( \gamma_0 \) 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 $\gamma_m$ values at $m=3$, 5, etc., which would be the FMI estimation when the default method was used. The $\gamma_{100}$ 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 $U_{100}$, respectively.

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 $\gamma_0$ estimate of the paper into a more correct $\gamma_0$ estimate.

### Table 2. Coefficient of variations (%) of $B_m$, $U_m$, and $\gamma_m$ for SIZE100

<table>
<thead>
<tr>
<th>$m$</th>
<th>MI-1, Anal-1</th>
<th>MI-2, Anal-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$B_m$</td>
<td>$U_m$</td>
</tr>
<tr>
<td>3</td>
<td>20.24</td>
<td>0.185</td>
</tr>
<tr>
<td>5</td>
<td>11.12</td>
<td>0.179</td>
</tr>
<tr>
<td>10</td>
<td>7.75</td>
<td>0.135</td>
</tr>
<tr>
<td>20</td>
<td>5.91</td>
<td>0.098</td>
</tr>
<tr>
<td>30</td>
<td>6.24</td>
<td>0.077</td>
</tr>
<tr>
<td>40</td>
<td>3.60</td>
<td>0.048</td>
</tr>
<tr>
<td>60</td>
<td>2.74</td>
<td>0.058</td>
</tr>
<tr>
<td>80</td>
<td>2.48</td>
<td>0.041</td>
</tr>
<tr>
<td>100</td>
<td>2.45</td>
<td>0.037</td>
</tr>
</tbody>
</table>

### Table 3. Comparison of different $\gamma_0$ 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

<table>
<thead>
<tr>
<th>$m$</th>
<th>Control $\hat{\gamma}_0 = \gamma_m$</th>
<th>Improved $\hat{\gamma}_0 = \gamma_m(m/(1 + m))$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\gamma}<em>0 = \gamma</em>{m \geq 100}$</td>
<td>$\hat{\gamma}_0 = c_m/(1 + c_m)$</td>
</tr>
<tr>
<td>3</td>
<td>0.00379</td>
<td>0.00283</td>
</tr>
<tr>
<td>5</td>
<td>0.00318</td>
<td>0.00265</td>
</tr>
<tr>
<td>10</td>
<td>0.00286</td>
<td>0.00260</td>
</tr>
<tr>
<td>20</td>
<td>0.00293</td>
<td>0.00279</td>
</tr>
<tr>
<td>30</td>
<td>0.00278</td>
<td>0.00269</td>
</tr>
<tr>
<td>40</td>
<td>0.00273</td>
<td>0.00267</td>
</tr>
<tr>
<td>60</td>
<td>0.00277</td>
<td>0.00267</td>
</tr>
<tr>
<td>80</td>
<td>0.00272</td>
<td>0.00269</td>
</tr>
<tr>
<td>100</td>
<td>0.00270</td>
<td>0.00267</td>
</tr>
<tr>
<td>(∞)</td>
<td>Best $\hat{\gamma}_0$: 0.00267</td>
<td></td>
</tr>
</tbody>
</table>
6. Conclusions
In most published researches, $\gamma_0$ and $\gamma_m$ are treated as synonyms. However, the two are different. The $\gamma_0$ is independent of MI, whereas $\gamma_m$ is a parameter of MI. $\gamma_m$ equals to $\gamma_0$ only if $B_0/U_0 = B_0/W_0$. To use MI for FMI estimation, one has to assume $\gamma_m = \gamma_0$, which will also be the assumption here.

The $\gamma_m$ decreases with the increase of $m$. We quantified the $m$–$\gamma_m$ relationship. The magnitude and the rate of the $\gamma_m$ decrease varies with $m$ and $\gamma_0$. At $m = 2$, $\gamma_2$ is greater than $\gamma_0$ by 50–81% depending on the $\gamma_0$ factor. At $m = 5$, the recommended $m$ value as being sufficient by some (e.g. Rubin, 1987), $\gamma_m$ is greater than $\gamma_0$ by 20–25% when $\gamma_0$ value ranges from 0.001 to 0.6. The decrease of $\gamma_m$ 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 $\gamma_m$ 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 $\gamma_m$ 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 $\gamma_m$ 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 $\gamma_m$ to $\gamma_0$ directly.

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

The current method $\gamma_0 = \gamma_m$ may result in a substantial FMI overestimation when $m$ is not sufficiently large. Three improved methods are proposed for estimating $\gamma_0$ from MI of a finite $m$. These three methods are (1) $\gamma_0 = \gamma_m_{0100}$, (2) $\gamma_0 = \gamma_m(m(m+1))$, and (3) $\gamma_0 = \gamma_m(c_m/c_m + 1)$, where $c_m = B_m/U_m$. In our MI trials, all three improved methods gave more accurate $\gamma_0$ estimates than $\gamma_0 = \gamma_m$ 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 $\gamma_0$. When $m$ is not sufficiently large, say, $m < 100$, the third method $\gamma_0 = \gamma_m(c_m/c_m + 1)$ should be one’s best option for $\gamma_0$ estimation. The second method $\gamma_0 = \gamma_m(m(m+1))$ has its value where $B_m$ and $U_m$ are not available and the only values available to use for $\gamma_0$ estimation are $m$ and $\gamma_m$.

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.

Author details
Qiyuan Pan
E-mail: qap1@cdc.gov
Rong Wei
E-mail: rw5@cdc.gov
1 National Center for Health Statistics (NCHS), 3311 Toledo Rd., Hyattsville, Maryland 20782, USA.
2 NCHS, 3311 Toledo Rd., Hyattsville, MD 20782, USA.

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

Citation information
Cite this article as: Improved methods for estimating fraction of missing information in multiple imputation, Qiyuan Pan & Rong Wei, Cogent Mathematics & Statistics (2018), 5: 1551504.

Note
1: 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


