December 2017

Missing Data in Longitudinal Surveys: A Comparison of Performance of Modern Techniques

Paola Zaninotto
University College London, p.zaninotto@ucl.ac.uk

Amanda Sacker
University College London, a.sacker@ucl.ac.uk

Recommended Citation

Missing Data in Longitudinal Surveys: A Comparison of Performance of Modern Techniques

Missing Data in Longitudinal Surveys: A Comparison of Performance of Modern Techniques

Paola Zaninotto
University College London
London, UK

Amanda Sacker
University College London
London, UK

Using a simulation study, the performance of complete case analysis, full information maximum likelihood, multivariate normal imputation, multiple imputation by chained equations and two-fold fully conditional specification to handle missing data were compared in longitudinal surveys with continuous and binary outcomes, missing covariates, and an interaction term.

Keywords: Chained equations, longitudinal data, maximum likelihood, missing data, random intercepts, two-fold fully conditional specification

Introduction

Full information maximum likelihood (FIML) estimation and multiple imputation (MI) are considered "state of the art" missing data techniques (Schafer & Graham, 2002, p. 147) and are highly recommended (Baraldi & Enders, 2010). They are superior to ad-hoc procedures, because they require less strict assumptions and yield unbiased estimates with missing completely at random (MCAR) and missing at random (MAR) missing data mechanisms. Recently, the use of FIML, multivariate normal imputation (MVNI), and imputation by chained equations (MICE) (also known as sequential regressions, regression switching, and fully conditional specification, FCS) has become increasingly popular. A number of comparisons using linear models, generalized models, and structural equation models were published (Collins, Schafer, & Kam, 2001; Newman, 2003; Acock, 2005; Ibrahim, Chen, Lipsitz, & Herring, 2005; van Buuren, 2007; Buhi, Goodson, & Neilands, 2008; Marshall, Altman, Royston, & Holder, 2010; Peyre, Leplege, & Coste, 2011; Ferro, 2014).

Dr Paola Zaninotto is a Senior Lecturer in Medical Statistics in the Department of Epidemiology and Public Health. Email them at: p.zaninotto@ucl.ac.uk.

The FIML approach has been widely used in structural equation modelling analysis; MICE and MVNI can be easily extended to repeated measures, and the two-fold fully conditional specification (two-fold FCS) was proposed as a special case of the fully conditional specification for repeated measures (Nevalainen, Kenward, & Virtanen, 2009). In the presence of many waves and many variables with missing data, MICE, as opposed to the two-fold FCS, can potentially cause numerical problems because of over-fitting and collinearity. The algorithm might not converge because of high correlations among variables with missing values. It might be required to reduce the categories of a discrete variable; furthermore, temporally-ordered data are not exploited. Two-fold FCS can handle imputation of missing data of both continuous and discrete variables in many waves. Each wave is imputed separately, using the past and/or future information for each of the variables to be imputed.

The comparative performance of these techniques simultaneously in a longitudinal setting remains unclear. This comparison is needed because an increasing number of longitudinal datasets are now becoming available, which are subject to missing data due to item non-response and attrition. The collected longitudinal data are therefore often incomplete with a non-monotone pattern (Minini & Chavance, 2004). MAR mechanisms and non-monotone patterns are common in most realistic settings (Horton & Kleinman, 2007). Good imputation techniques for dealing with missing values in longitudinal data structured in a long format (1 record per observation-wave) are still missing. FIML and the MI techniques (MVNI, MICE, and two-fold FCS) can create valid multiple imputations under MAR (Enders, 2001a; van Buuren, 2007; Schafer & Graham, 2002; Nevalainen et al., 2009) and for monotone and non-monotone patterns, when data are structured in a wide format (i.e. unstacked data, one observation row per each subject with the measures for each occasion in separate columns). Data can then be converted into a long format file for analysis.

The aim of this study it so investigate the usefulness of the ML and MI methods in the context of longitudinal analysis with a large number of missing data points. The realistic situation of missing data in more than one outcome is considered, as well as accommodation of an interaction term and data missing on more than one independent variable. Generally, analyses should exclude individuals with imputed data on the outcome (von Hippel, 2007). However, in settings where one outcome is also the exposure for the other outcome, excluding individuals with imputed data may result in selection bias (Young & Johnson, 2010). Therefore, it was suggested that including imputed data on the outcome is feasible when auxiliary variables (or as in our case outcome variables) are highly correlated

with the outcome (White, Royston, & Wood, 2011). Hence, the purpose of this study is to explore the performance of the missing data techniques in recovering the true parameters estimated from two random intercepts models, one with a continuous outcome and the other with a binary outcome. A simulation study is set up, where FIML is compared to three MI techniques: multivariate normal imputation (MVNI), multiple imputations by chained equations (MICE) and the recently proposed two-fold fully conditional specification (two-fold FCS). The missingness pattern is non-monotone and the missing data mechanism is missing at random (MAR). We show results of complete case analysis because it is an approach commonly used.

Methodology

Empirical Data

The aim of the simulation study was to evaluate the four techniques for handling missing data (FIML, MVNI, MICE, and two-fold FCS) in a longitudinal survey of ageing to explore gender differences in trajectories of quality of life and depression among people with coronary heart disease (CHD), adjusting for covariates. For this purpose the first three waves (2002-03 to 2006-07) of the English Longitudinal Study of Ageing, described elsewhere (Steptoe, Breeze, Banks, & Nazroo, 2013) was used. ELSA was established in 2002-2003 (wave 1), is a biannual, ongoing, nationally representative, prospective cohort study of people aged 50 years and over living in private households in England. In the first wave 11,391 respondents were interviewed.

The analytical sample of this study consisted of 4,496 participants in wave 1 with CHD and healthy participants (without CHD and known longstanding conditions at baseline); 3,465 in wave 2 (2004-05); and 3,031 in wave 3 (2006-07). A total of 1,998 participants had complete data on all variables across the three waves. The data consisted of two completely observed exposures (CHD and sex) and their interaction (CHD*sex), and two incomplete dependent variables: Quality of life (QoL) score ranging from 0 to 57 (approximately normally distributed), with higher scores indicating better quality of life; and a binary variable measuring depression ("0" no; "1" yes). The correlation between quality of life and depression is -0.45. The following covariates were also included: age (completely observed, normally distributed); marital status (categorical, three categories); wealth (categorical, three categories); physical activity (categorical, three categories); smoking status (categorical, three categories); and alcohol consumption

(categorical, three categories). With the exception of sex, all variables are time-varying. The ELSA dataset is publicly available via registration with the UK data service (https://www.ukdataservice.ac.uk/).

Obtaining the True Parameters from the Empirical Data

Random intercepts models (Goldstein, 2003) were used to explore gender differences in quality of life and depression among older people with coronary heart disease (CHD), adjusting for covariates. The 1,998 individuals with complete data were treated as if they were the underlying population, and the true parameters were obtained from the two random intercepts models estimated using the complete data. The first model is for the continuous outcome and was estimated as follows:

$$y_{ij} = \beta_0 + \sum_{p=1}^{3} \beta_p x_{pj} + \sum_{p=1}^{8} \vartheta_p x_{pij} + u_j + e_{ij} \tag{1}$$

where y_{ij} is the quality-of-life for individual j at time i, p denotes the number of time-varying and time invariant variables. x_{pj} are the time-invariant factors ($p = 3$) gender, CHD (at wave 1), and the interaction term between CHD and gender; x_{pij} are the time-varying factors ($p = 8$) age (a linear and quadratic term), cohabitation status, depression, wealth, smoking status, alcohol consumption and physical activity. u_j denotes the random effect accounting for the individual level variation and e_{ij} is a combination of random variation and measurement error specific to each occasion i for an individual j. The random effects have variances equal to σ_u^2 and σ_e^2 at the individual and occasion levels. The true parameters, obtained from (1) on the complete data, are: coefficient for CHD $\beta_1 = -1.64$ (s.e. 0.47); coefficient for sex $\beta_2 = 0.84$ (s.e. 0.28), coefficient for the interaction term between CHD and sex $\beta_3 = -0.19$ (s.e. 0.71), $\sigma_u^2 = 24.2$ (s.e. 1.10), $\sigma_e^2 = 24.8$ (s.e. 0.57), adjusted for the covariates described above (true parameters for the covariates are presented in supplemental resources).

A logit model was estimated for the binary outcome as follows:

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + \sum_{p=1}^{3} \beta_p x_{pj} + \sum_{p=1}^{6} \beta_p x_{pij} + u_j \tag{2}$$

where

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$$

is the log-odds that $y_{ij} = 1$ (i.e. the probability of having depression) at occasion i for individual j. All the other variables are the same as in model (1) with the exception of depressive symptoms and the quadratic term for age, which are not included as covariates; therefore $p = 6$ in the second part of the equation, u_j denotes the random individual level effect with variance σ_u^2. The true parameters obtained from (2) on the complete data are: coefficient for CHD $\beta_1 = 0.65$ (s.e. 0.25); coefficient for sex $\beta_2 = 0.74$ (s.e. 0.16), coefficient for the interaction term between CHD and sex $\beta_3 = 0.03$ (s.e. 0.36), $\sigma_u^2 = 4.02$ (s.e. 0.44), adjusted for the covariates described above, (true parameters for the covariates are presented in online resources).

Data Simulation

We used a non-parametric simulation study based on the observed real data consisting of 1,998 ELSA participants with complete data at all three waves. Starting with the data of complete observations, missingness was imposed on the two dependent variables (quality of life and depression) and on five covariates (marital status, wealth, physical activity, smoking status and alcohol consumption). Generation of missing data was performed using the data in wide format, in which there is one observation row for each subject with each measurement represented as a different variable, therefore each time-varying variable was present three times for each subject. Random uniform numbers were used to reproduce the same probabilities of missingness as those occurring in the real data: if the rank of the random number was equal to or less than the probability observed, a missing value was generated for the variable of interest. Variables from the first wave only required generation of missing data due to item non-response, variables from waves 2 and 3 required the generation of missing data due to attrition in addition to item non-response, which were performed separately according to the specific probabilities observed in the real data. In order to set a MAR scenario, within each wave, missingness probabilities were allowed to depend on observed variables in the dataset, independently of the missing values themselves. Deletion was repeated for each of the 1,000 data sets. The resulting amount of missingness ranged between 54% and 57% (resulting in the sample size of each replicate ranging from 859 to 919).

Missing Data Methods for the Simulated Data

Missing data on the 1,000 data sets were treated using the following methods: FIML, MVNI, MICE, and two-fold FCS. For the latter three MI techniques, five data sets for each replicate were imputed. Most literature (Rubin, 1987; van Buuren et al., 1999) suggests that good inferences can be made with the number of imputed data sets (m) as few as $m = 5$. Rubin (1987) showed the efficiency of an estimate based on m imputations, relative to one based on an infinite number, is $(1 + \lambda/m)^{-1}$, where λ is the rate of missing information. In this setting, with approximately 56% missing information, $m = 5$ imputations is 90% efficient. Complete case analysis (average sample size across the simulated data 887) was also conducted. All analyses were carried out in Stata version 12

Full Information Maximum Likelihood (FIML)

The FIML technique does not impute, or fill in missing values, but directly estimates model parameters and standard errors using all available raw data (Enders, 2001a). The FIML estimator maximizes a likelihood function that is the sum of n casewise likelihood functions (where n is the number of respondents). As noted by Enders (2001a), assuming multivariate normality, the following function is maximized:

$$\log L_i = K_i - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_i)^\mathrm{T}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i) \tag{3}$$

such that $\mathbf{x}_i$ is the vector of complete data for case i, $\boldsymbol{\mu}_i$ is the vector of mean estimates for those variables that are observed for case i, and K_i is a constant that depends on the number of complete points for case i. The determinant and inverse of the covariance matrix $\boldsymbol{\Sigma}_i$ are based only on those variables that are observed for case i. This likelihood function measures the discrepancy between the observed data and current parameter estimates using all available data for a given case.

Summing over the n casewise functions yields the discrepancy function for the entire sample:

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n}\log L_i \tag{4}$$

To illustrate how FIML works, consider a model with four observed variables: X_1, X_2, X_3, and X_4. The parameters of interest are

$$\mathbf{\mu} = [\mu_1, \mu_2, \mu_3, \mu_4] \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

The likelihood value for a subject with missing X_1 would be a function of the values on the observations for the other three variables, X_2, X_3, and X_4, as well as the parameter estimates that involved these three variables. The relevant parameters are shown in the following:

$$\mathbf{\mu} = [., \mu_2, \mu_3, \mu_4] \quad \text{and} \quad \Sigma = \begin{bmatrix} . & . & . & . \\ . & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ . & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ . & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

By contrast, the likelihood value for a subject with missing X_2 and X_4 would be a function of the two other observations (X_1 and X_3) as well as the parameter estimates that involved X_1 and X_3. The relevant parameters are shown in the following:

$$\mathbf{\mu} = [\mu_1, ., \mu_3, .] \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_{11} & . & \sigma_{13} & . \\ . & . & . & . \\ \sigma_{31} & . & \sigma_{33} & . \\ . & . & . & . \end{bmatrix}$$

Then the value of the overall discrepancy function is obtained by summing the likelihood functions for each individual.

Enders (2001a) explained that, at a more conceptual level, it is assumed that missing values on a variable X are conditionally dependent on other variables in the data (missing at random, MAR), and incorporating vectors of partially complete data in the individual level likelihood functions (3) implies probable values for the missing data during the parameter estimation process. Conceptually this is analogous to generating predicted scores for the missing data by regressing X on

other variables used in the analysis. Although the method assumes multivariate normality, it has been shown that it can also be extended to situations in which multivariate normality does not hold (Enders, 2001b).

Multivariate Normal Imputation (MVNI)

In a multivariate normal imputation model, missing data are imputed using simultaneous linear regression models in which each variable potentially depends on all other variables (Schafer & Olsen, 1998). MVNI assumes a joint multivariate normal distribution for all variables. When dealing with an arbitrary missing data pattern it is possible to use data augmentation (DA) to generate imputed values assuming an underlying multivariate normal distribution. DA is an iterative Markov Chain Monte Carlo (MCMC) procedure which is suitable for arbitrary missing data patterns. Suppose $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ are multivariate data decomposed into an observed part $\mathbf{Y}_{obs}$ and a missing part $\mathbf{Y}_{mis}$ from a normal distribution $P(\mathbf{Y} \mid \boldsymbol{\theta}) = N(\beta, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. In many incomplete-data problems, the observed-data posterior $P(\boldsymbol{\theta} \mid \mathbf{Y}_{obs})$ is intractable and cannot be simulated easily. When $\mathbf{Y}_{obs}$ is augmented by an assumed latent value of the $\mathbf{Y}_{mis}$, the resulting conditional posterior distribution $P(\boldsymbol{\theta} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ becomes much easier to handle. DA for missing data consists of two steps, the Imputation step (I-step) and the Posterior step (P-step). In the I-step, given a current $\boldsymbol{\theta}^{(t)}$, a value for the missing data-point is first drawn from the conditional predictive distribution of $\mathbf{Y}_{mis}^{(t+1)} \sim P\left(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \boldsymbol{\theta}^{(t)}\right)$ given the observed data $\boldsymbol{\theta}^{(t)}$. In the P-step, a new value of $\boldsymbol{\theta}^{(t+1)}$ is drawn from its conditional posterior $P\left(\boldsymbol{\theta} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(t+1)}\right)$ given $\mathbf{Y}_{mis}^{(t+1)}$. These new estimates are used in the next I-step. Without prior information about the parameters a non-informative prior distribution is used. Iterating these two steps creates a Markov chain of length $t\left\{\left(\mathbf{Y}_{mis}^{(t)}, \boldsymbol{\theta}^{(t)} : t = 1, 2, \ldots\right)\right\}$ which converges to $P(\boldsymbol{\theta}, \mathbf{Y}_{mis} \mid \mathbf{Y}_{obs})$. The two steps are iterated long enough for the results to be reliable for a multiply imputed data set.

Here, the posterior mode, the highest observed-data posterior density with a non-informative prior, is computed from the Expectation Maximization (EM) algorithm, and is used as the starting value for the chain. The MI procedure takes 200 burn-in iterations before the first imputation and 100 iterations between imputations. In a Markov chain, the information in the current iteration has influence on the state of the next iteration. The burn-in iterations are iterations at the beginning of each chain that are used to eliminate the dependence on the starting

value of the chain and to achieve a stationary distribution. Although the regression and MCMC methods assume multivariate normality, inferences based on multiple imputation can be robust to departures from the multivariate normality assumption if the amount of missing information is not large. It makes sense to use a normal model to create multiple imputations even when the observed data are somewhat non-normal, as supported by simulation studies described in Schafer (1997) and the original references therein. The imputation model included the same variables as the substantive models including the interaction term between CHD and gender and the linear and quadratic effects of age (which were completely observed).

Multiple Imputation by Chained Equations (MICE)

MICE does not start with the construction of a well-defined joint distribution for the variables to be imputed. Instead, it starts with a collection of univariate conditional distributions for variables with missing data in terms of all other variables. The main idea is that a sequence of univariate conditional models is constructed for each potentially missing variable (dependent and/or explanatory) with fully conditional specifications of prediction equations. The other variables (potentially missing or complete) are used as explanatory variables in each univariate imputation model. The standard procedure for creating multiple imputations of each potential missing variable Y_1, Y_2,..., Y_p using complete predictors $\mathbf{X}$ as independent variables for the tth iteration of the method is described as follows:

$$
\begin{aligned}
Y_1^{(t+1)} &\sim \mathrm{f}_1\left(Y_1 \mid Y_2^{(t)},\ldots,Y_p^{(t)},\mathbf{X},\theta_1\right) \\
Y_2^{(t+1)} &\sim \mathrm{f}_2\left(Y_2 \mid Y_1^{(t+1)},Y_3^{(t)},\ldots,Y_p^{(t)},\mathbf{X},\theta_2\right) \\
&\vdots \\
Y_p^{(t+1)} &\sim \mathrm{f}_p\left(Y_p \mid Y_1^{(t+1)},Y_2^{(t+1)},\ldots,Y_{p-1}^{(t+1)},\mathbf{X},\theta_p\right)
\end{aligned}
\tag{5}
$$

Imputed values are drawn from (5) for iterations $t = 0$, 1,..., T until convergence is reached at $t = T$. The θ_1,..., θ_p are the model parameters with a uniform prior, the univariate imputation models f_1,..., f_p are chosen to be appropriate for imputing Y_1,..., Y_p. This means that logistic regression can be used for binary variables, linear regression for continuous, ordinal logistic regression for categorical variables and so forth (van Buuren, 2007). Univariate posterior draws are made one variable at a time by cycling through all p models given current values

of the other variables (Molenberghs & Kenward, 2007). After sufficient cycles (10-20), the imputations are taken from one final cycle through the univariate model. The ordering of univariate imputations is not relevant.

MICE differs from the MVNI in that it does not start with the construction of a well-defined joint distribution for the variables to be imputed.

The imputation model included the same variables as the substantive models including the interaction term between CHD and gender and the linear and quadratic effects of age (which are completely observed). For the variable quality of life we used predictive mean matching (Royston & White, 2011), thus constraining imputed values to the set of observed values. We number of cycles used in the imputation was 100.

Two-fold Fully Conditional Specification (FCS)

Two-fold FCS (Nevalainen et al., 2009) is a special case of MICE described above. MICE for Y_l ($l = 1, \ldots, p$) is extended to q repeated waves, as follows:

$$\mathrm{f}\left(\mathbf{Y}_u^{\mathrm{mis}} \mid Y_{u-1}^*, Y_{u+1}^*, X_{u1}, \ldots, Z_{up}\right), \quad u = 1, \ldots, q$$

At time i, Y_u is imputed conditional on the same variable observed at time $u - 1$ and $u + 1$, and the other p variables Z at time u. Similarly, at time u, Z_u with missing data is imputed conditional on the same variable observed at time $u - 1$ and $u + 1$, the other p variables Z, and the Y. One iteration runs over the variables $l = 1, \ldots, p$, called within-time iteration. The past and future observations (Y_{u-1} and Y_{u+1}) are not imputed at this stage, they serve only in the role of predictors in the imputation model. There is also a second imputation iteration over waves ($u = 1, \ldots, q$), called among-time iteration. The method is fully described in Nevalainen et al.

Two-fold FCS differs from MICE in that each time point is imputed separately. The method uses information recorded before and after the time point to impute missing data. The past and future values might have missing data which are imputed by default, but are only used as predictors at that given time.

The application of two-fold FCS to the simulation data was performed as follows and described graphically in Figure 1:

1) Variables with missing data at wave 1 were imputed using as predictors all other variables in the same wave, plus the future observation (at wave 2) of the same variable. For example, to impute

missing values for the continuous outcome (QoL) at wave 1, the following linear regression model is used:

$$QoL_{wave\ 1} = Sex_{wave\ 1} + CHD_{wave\ 1} + (Sex*CHD)_{wave\ 1}$$
$$+ Age_{wave\ 1} + Age^2_{wave\ 1} + Marital_{wave\ 1}$$
$$+ Wealth_{wave\ 1} + Smoking_{wave\ 1} + Physicalact_{wave\ 1}$$
$$+ Alcohol_{wave\ 1} + Depression_{wave\ 1} + QoL_{wave\ 2}$$

In this example the variable QoL at wave 2 is used as a predictor to impute missing data for QoL at wave 1. Because QoL at wave 2 has missing data as well, MICE by default imputes missing data for that variable too. However, the imputed values of QoL at wave 2 are then dropped from the imputed data set as, in the imputation of wave 1, it serves only the role of a predictor and it will be imputed in the next stage.
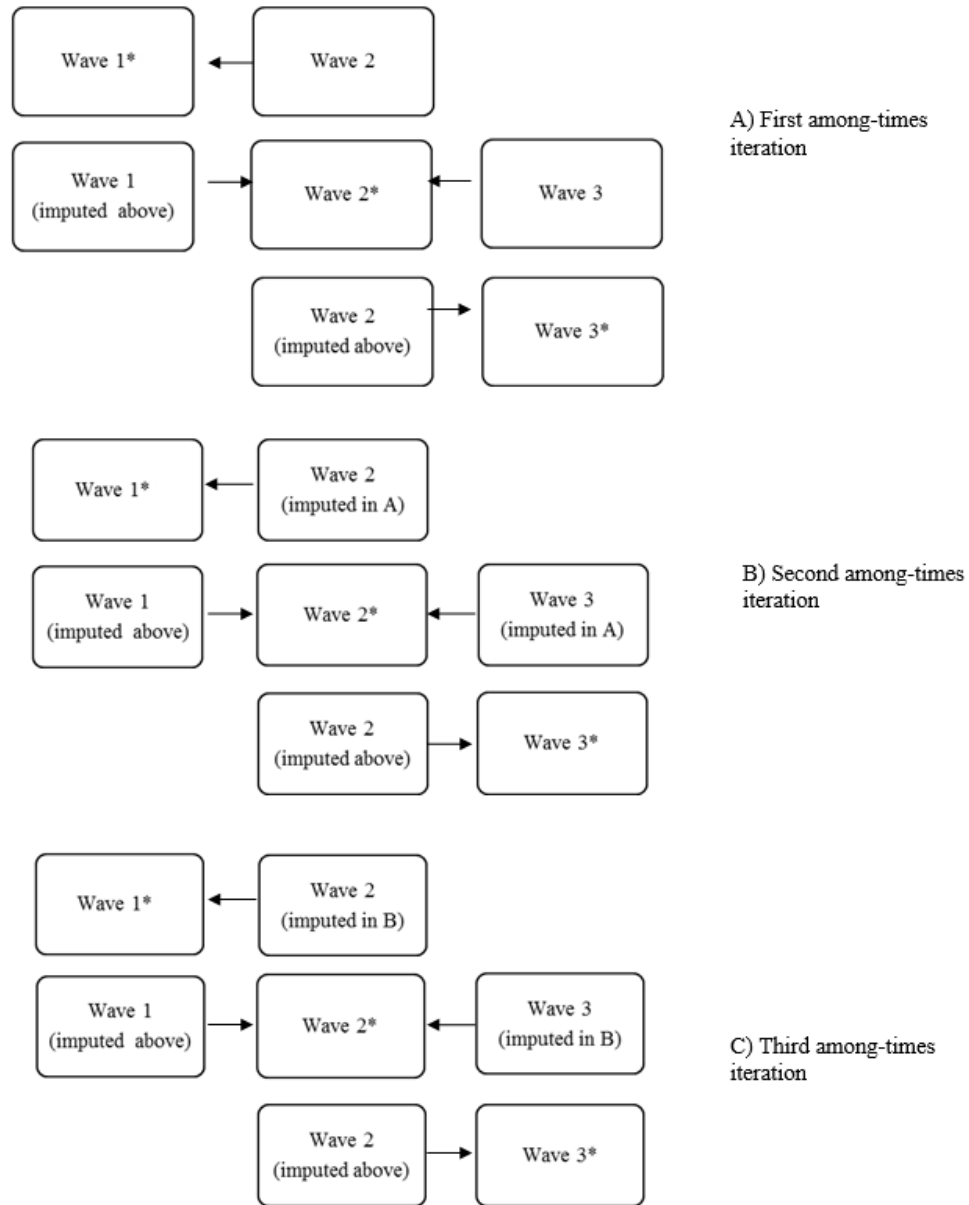
2) Variables with missing data at wave 2 are imputed using as predictors all other variables in the same wave, plus the past (wave 1 including values imputed in the previous step) and future (wave 3) observations of the same variable. The imputed values of these variables from wave 3 are then dropped after the imputation as at this stage they serve only the role of predictors in the imputation model.

3) Variables with missing data at wave 3 are imputed using as predictors all other variables in the same wave, plus the past observations (wave 2 imputed in the previous step) of each variable to be imputed.

Figure 1 gives a graphical explanation of steps 1), 2), and 3).

Steps 1) to 3) form one among-times iteration. It must be decided how many among-times iterations are needed. Nevalainen et al. (2009) showed that increasing the number of iterations from one to five improved the performance of the estimators although the gain due to the increase was relatively small. We used three among-times iterations and compared the means of the imputed variables at each wave with the means of the complete case data. It was decided that three among-times iterations were enough because the estimates from the five imputed data sets were very close to those of the underlying population.

To summarize, steps 1) to 3) were repeated 3 times (three among-times iterations) to generate 1 imputed dataset, the procedure was then repeated four more times to obtain five imputed datasets each for the 1000 data sets. The imputation

model included the same variables as the substantive models including the interaction term between CHD and gender and the linear and quadratic effects of age (which are completely observed).

Figure 1. Two-fold fully conditional specification; * indicates the variable with missing data at the specific wave that is to be imputed

Data Analysis of Simulated Data

After the imputation stage had been completed on the 1,000 data sets for each MI technique (MVNI, MICE, and two-fold FCS), analyses based on the imputed data sets were carried out in Stata version 12. Linear random intercepts models were estimated for the continuous outcome (QoL) and random intercepts logit models were estimated for the binary outcome (depression), described earlier. The estimates from the analysis were stored. From the stored estimates, some summary measures (Burton, Altman, Royston, & Holder, 2006) were calculated to assess each missing data strategy as follows:

The (average) estimate of interest:

$$\bar{\beta} = \frac{\sum_{k=1}^{r} \hat{\beta}_k}{n}$$

where r is the number of data sets (1,000), and $\hat{\beta}_k$ is the estimate of interest within each of the $k = 1,\dots, r$ data sets. When MI is performed, each $\hat{\beta}_k$ is the overall estimate obtained according to Rubin's formula (Rubin, 1987), which is just the average of the 5 combined estimates within each of the $k = 1,\dots, r$ data sets.

The (average) standard error of the estimate of interest:

$$\text{SE}\left(\hat{\beta}\right) = \frac{\sum_{k=1}^{r} \text{SE}\left(\hat{\beta}_k\right)}{n}$$

where $\text{SE}\left(\hat{\beta}_k\right)$ is the standard error of the estimate of interest within each of the $k = 1,\dots, r$ replications. When MI is performed, each $\text{SE}\left(\hat{\beta}_k\right)$ is the overall standard error of the estimate of interest obtained from the five combined estimates according to Rubin's formula (Rubin, 1987), within each of the $k = 1,\dots, r$ data sets.

Evaluation Criterion

In order to evaluate the performance of each procedure employed to deal with missing data, we used assessments of accuracy and precision. Accuracy indicates the degree of closeness of the estimated value to the true parameter; precision refers to the repeatability or reproducibility of the measurement.

For the assessment of accuracy the following were used:

Bias: $\left(\bar{\hat{\beta}} - \beta\right)$ which is the difference between the average estimate and the true parameter (Schafer & Graham, 2002; Sinharay, Stern, & Russell, 2001).

The Mean Square Error (MSE): $\left(\bar{\hat{\beta}} - \beta\right)^2 + \left(\text{SD}\left(\hat{\beta}\right)\right)^2$ is the average squared difference between the estimate and its true parameter plus its variance; therefore it can be seen as a summary of both bias and variability. A value of the MSE close to zero indicates that the average estimator predicts the true parameter with good accuracy.

For the assessment of precision the following was used:

Standardized bias percent:

$$100 \times \frac{\left(\bar{\hat{\beta}} - \beta\right)}{\text{SE}\left(\hat{\beta}\right)}$$

which is the bias as a percentage of the standard error. A standardized bias is considered to have a large impact on the precision if its absolute value exceeds 40 per cent (Collins et al., 2001).

Results

The results of the comparison of the missing data techniques for the analysis of the continuous outcome (quality of life) are shown in Table 1. The bias of the coefficients for CHD and sex are small and close to zero for the three MI techniques, which seem to give estimates of the parameter for sex closer to the true parameter compared to FIML. The largest bias in the coefficient for the CHD by sex interaction term was obtained under FIML and least bias was obtained under MVNI and MICE, although the values of the MSE were similar for all techniques. The MSE is the sum of the squared bias and the variance, therefore it can be evinced that the variability across replicates is lower for each of the three parameters estimated under FIML than under the three MI methods.

The three MI techniques produced a more precise estimate of the coefficient for the interaction term than FIML, as shown by the smaller values of the Standardized bias percent. This is probably due to the fact that imputation using MICE, two-fold FCS, and MVNI models included the interaction term and in that sense they reflected the substantive model. Estimates of the between (i.e. individual) variance obtained from MVNI and MICE were closer to the true parameter compared to the estimates obtained from the other methods. The estimates of the within (i.e. wave specific) variance were close to the true parameter for all methods, except for the estimates produced by the two-fold FCS. Estimates obtained from complete case analysis (average sample size across simulated data 887) showed good precision but did not achieve good levels of accuracy. Results comparing the true parameters with each method for missing data for the time-varying variables are presented in supplemental Tables S1 to S5. All methods performed equally well in producing estimates close to the true parameters for most of the time-varying covariates. MVNI and Two-fold FCS produced estimates slightly less precise the true parameters of cohabiting status and wealth. MICE outperformed the other methods in producing estimates that were close to the true parameter of depression.

Reported in Table 2 are the results of the comparison of the missing data techniques for the analysis of the binary outcome (depression). The large values of the biases and of the standardized bias percent suggest that FIML did not produce estimates that were close to the true parameters for sex, CHD, the interaction term between sex, and CHD and the between variance. Whereas the estimates produces by MICE were very close to the true coefficients for sex and CHD, and overall showed good levels of precision and accuracy as demonstrated by the MSE and standardized bias percent. Estimates of the interaction term between sex and CHD obtained from MICE and MVNI were slightly less close to the true parameter compared to the estimated obtained from two-fold FCS. Estimates of the coefficients for sex and CHD obtained from MVNI were not as close to the true parameters as those obtained from MICE and two-fold FCS. The estimate of the between variance obtained from MICE showed good accuracy and precision. Estimates obtained under complete case analysis (average sample size across simulated data 887) showed good accuracy as evinced from the small bias and good precision according to the values of the standardized bias percent. Results comparing the true estimates for the time-varying variables and each method for missing data are presented in supplemental Tables S6 to S10. Estimates obtained from all methods were close to the true parameters and showed good accuracy and precision.

Table 1. Summary of validity criteria for true parameters compared to the methods for missing data applied to the model of the continuous outcome quality of life

		$\boldsymbol{\beta}$
True parameters	CHD	-1.64
	Sex	0.84
	CHD*Sex	-0.19
	Between variance	24.21
	Within variance	24.79

		$\overline{\hat{\boldsymbol{\beta}}}$	$\mathbf{SE}(\hat{\boldsymbol{\beta}})$	**Bias**	**Stand. Bias %**	**MSE**
Complete Case	CHD	-1.60	0.70	0.70	6.7	0.321
	Sex	0.87	0.42	0.42	6.4	0.104
	CHD*Sex	-0.26	1.06	1.06	-6.9	0.600
	Between variance	24.80	1.03	1.03	0.8	1.053
	Within variance	24.01	1.07	1.07	-18.0	1.724
FIML	CHD	-1.54	0.48	0.10	20.8	0.041
	Sex	1.14	0.29	0.30	104.3	0.015
	CHD*Sex	-0.54	0.72	-0.35	-48.5	0.093
	Between variance	23.09	1.22	-1.11	91.2	1.861
	Within variance	25.05	0.72	0.25	-34.9	0.538
MVNI	CHD	-1.59	0.51	0.05	10.1	0.046
	Sex	0.79	0.31	-0.05	-14.8	0.019
	CHD*Sex	-0.24	0.77	-0.05	-6.7	0.104
	Between variance	23.59	1.06	-0.61	-57.9	1.492
	Within variance	24.85	1.03	0.05	5.1	1.169
MICE	CHD	-1.64	0.52	0.00	0.7	0.049
	Sex	0.86	0.31	0.02	6.1	0.019
	CHD*Sex	-0.22	0.79	-0.03	-4.4	0.115
	Between variance	24.33	1.06	0.13	12.1	0.818
	Within variance	25.01	1.03	0.21	20.7	0.556
Two-fold FCS	CHD	-1.63	0.51	0.01	1.7	0.066
	Sex	0.87	0.31	0.03	10.7	0.023
	CHD*Sex	-0.24	0.77	-0.06	-7.2	0.123
	Between variance	22.79	1.06	-1.41	-133.4	2.787
	Within variance	25.85	1.03	1.05	101.9	1.628

Note: Estimates adjusted for age, age squared, marital status, wealth, physical activity, smoking, alcohol consumption, depression, for which missing values have been imputed. CHD=Coronary heart disease. FIML=Full information maximum likelihood. MVNI= Multivariate normal imputation. MICE=Multiple imputation by chained equation. FCS=Fully conditional specification. SE=Standard error. MSE=Mean square error.

Table 2. Summary of validity criteria for true parameters compared to the methods for missing data applied to the model of the binary outcome depression

		β
True parameters	CHD	0.65
	Sex	0.74
	CHD*Sex	0.03
	Between variance	4.02

		$\bar{\hat{\beta}}$	$SE(\hat{\beta})$	Bias	Stand. Bias %	MSE
Complete Case	CHD	0.64	0.38	-0.01	-3.0	0.079
	Sex	0.75	0.24	0.01	3.7	0.030
	CHD*Sex	0.03	0.55	0.00	0.6	0.160
	Between variance	3.98	1.18	-0.04	-3.8	0.227
FIML	CHD	0.44	0.24	-0.21	-90.1	0.061
	Sex	0.50	0.14	-0.24	-176.3	0.067
	CHD*Sex	0.20	0.34	0.18	51.7	0.048
	Between variance	2.94	0.16	-1.08	-675.0	1.526
MVNI	CHD	0.58	0.26	-0.07	-25.7	0.015
	Sex	0.65	0.16	-0.09	-53.6	0.012
	CHD*Sex	0.04	0.37	0.01	3.7	0.018
	Between variance	3.47	1.14	-0.55	-48.4	0.381
MICE	CHD	0.66	0.28	0.01	2.3	0.013
	Sex	0.74	0.17	0.00	-1.4	0.005
	CHD*Sex	0.01	0.39	-0.02	-6.2	0.018
	Between variance	3.86	1.14	-0.16	-14.1	0.099
Two-fold FCS	CHD	0.62	0.26	-0.03	-12.5	0.013
	Sex	0.70	0.16	-0.04	-22.4	0.007
	CHD*Sex	0.03	0.37	0.00	-0.4	0.024
	Between variance	3.40	1.14	-0.62	-54.7	0.465

Note: Estimates adjusted for age, marital status, wealth, physical activity, smoking, alcohol consumption, for which missing values have been imputed. CHD=Coronary heart disease. FIML=Full information maximum likelihood. MVNI= Multivariate normal imputation. MICE=Multiple imputation by chained equation. FCS=Fully conditional specification. SE=Standard error. MSE=Mean square error

Conclusion

This simulation study was based on a large, national longitudinal survey to assess the problem of handling a non-monotone pattern of missing data. The aim of this study was to obtain valid and efficient estimates of regression coefficients from random intercepts models fitted to longitudinal data (three waves). The data set for

this study had incomplete time-dependent outcomes (one continuous and one binary) and time-dependent and time-independent covariates (of different types), therefore it was necessary to accommodate missingness within each wave, as well as unit non-response at all follow-up waves. In order to investigate which technique could be suitable with this structure of the data, FIML was compared with three MI techniques: MVNI, MICE and the recently proposed two-fold FCS. The performance of each of the methods appeared to vary according to the type of outcome.

The continuous outcome variable was the variable with the largest proportion of missing data, and yet all four missing data techniques performed well, although the MI techniques showed better accuracy and precision than FIML. Complete case analysis did not achieve good levels of accuracy. Additionally, the three MI techniques produced estimates the interaction term that showed better precision and smaller bias values compared to the estimate obtained from FIML. This is an advantage of multiple imputation techniques: the interaction term can and should be accommodated in the imputation model thus reflecting the substantive model. MICE and MVNI produced estimates of the within (wave) variance close to the true estimates, although all methods except two-fold FCS produced good estimates of the between (individual) variance.

A different picture was given by the results involving the binary outcome (depression). Estimates obtained under complete case analysis achieved good levels of precision and accuracy, according to the values of the bias. Not surprisingly the methods that assume a joint normal distribution for the non-normally distributed outcome FIML and MVNI did not produce estimates that had the same level of bias and precision as those obtained with more flexible chained equations methods. Levels of accuracy and precision were less good for the estimates obtained under FIML. Also, estimates of the MVNI showed less precision compared to other MI methods. However, estimates obtained from MICE and two-fold FCS, compared to both FIML and MVNI, showed better accuracy and precision. The flexibility of MICE and two-fold FCS becomes obvious in the presence of a non-normal outcome for which an appropriate conditional distribution (logistic) was specified in the imputation stage. The estimates of the between variance of the binary outcome obtained from MICE were slightly closer to the true parameters and showed better precision than those obtained from two-fold FCS.

FIML is relatively easy to implement and it performed almost as well as MI techniques with the continuous outcome. However, the estimate of the interaction term was less precise and not as close to the true parameters as those obtained from the MI techniques.

In general, good estimates in terms of accuracy and precision were obtained from MVNI for the continuous outcome. However, less precise and accurate estimates were obtained from MVNI in the case of the binary outcome. This is not surprising, given that the amount of missing data was relatively large (between 54% and 57%). It has been suggested that inferences based on multiple imputation can be robust to departures from the multivariate normality assumption if the amount of missing information is below 50% (Schafer, 1997).

The advantage of MICE and two-fold FCS is that each type of variable (continuous, binary, unordered, and ordered categorical) is modelled separately (Molenberghs & Kenward, 2007). For both outcomes, MICE showed slightly better precision and accuracy than two-fold FCS and the other two techniques.

The MICE procedure used all the longitudinal data to estimate missingness whereas two-fold FCS imputed each time point separately and used only information on the same variable recorded prior to and post a given time point. Therefore it was anticipated that estimating missingness using MICE would result in a slightly better performance.

In the simulation study consisting of three waves, estimates obtained from MICE were in general more accurate and precise; however, when the longitudinal data to be imputed have many time points (waves) per observation, using MICE may not be computationally feasible, especially for discrete variables. Researchers may opt for two-fold FCS which uses a doubly-iterative procedure: each wave is imputed separately using only the relevant past and future observations, therefore the intensiveness of the computation is reduced but the benefit of using past and future observation is retained.

Another approach to impute longitudinal data with many waves per observation is to structure the data in a long format and perform imputation in MICE ignoring the dependence of data across waves. Although this method is conceptually wrong, it is appealing to many researchers. We anticipate that data structured in long format and imputed with MICE would perform more poorly than two-fold FCS. The main reason is that ignoring the dependence between observations might result in inefficient estimates of the parameters of interest. Two-fold FCS on the other hand accounts for the correlation between individual's responses over time and therefore should (and in our simulation did) recover the parameters of interest well. Further studies are needed to test the performance of two-fold FCS, especially in the presence of many follow-up points, when the implementation of MICE is not feasible.

A major strength of this study is the use of a real data set to provide a suitable structure for simulating the 1,000 data sets, which simplifies the data generation

procedures and avoids arbitrary choices. Also, by replicating the patterns of missingness seen in the original incomplete data set, a realistic framework was provided for simulating the missing data (Marshall et al., 2010). Another major strength is that the recently proposed two-fold FCS was applied to longitudinal data from a national survey and compared with FIML and MVNI. Previously, only FIML with MVNI and FCS was compared with cross-sectional data. Also, the data time-invariant variables were completely observed. Nevertheless, the MI methods used are suitable for the imputation of missing data in time-invariant variables too. Time-invariant variables only appear once for each individual because the data are structured in a wide format. Therefore, strength of imputing longitudinal data in wide format is that imputed values of time invariant variables cannot end out time varying.

One of the possible limitations of this simulation study is that missing data were generated under MAR. The plausibility of the MAR assumption could have been affected by the fact that auxiliary variables were by design not included in the imputation model. In analysis not shown here, auxiliary variables found to be associated with the values of variables with missing data (such as housing tenure, number of people in the household, and Government Office Region) were used to impute missing data using the three MI techniques. The ability to recover the true parameters by MVNI, MICE, and two-fold FCS techniques did not depend upon the addition of these variables; rather they helped reinforce the MAR assumption. However, it was necessary to exclude auxiliary variables in order to make the four techniques comparable. It may be appropriate to strengthen the MAR assumptions and reduce the chance obtaining multiple imputations with similar bias and precision to a complete case analysis may decide to opt for one of the MI techniques presented here rather than FIML which does not allow the inclusion of auxiliary variables.

Auxiliary variables will also reduce bias compared to complete case analysis. Although it is generally argued that analyses should exclude individuals with imputed data on the outcome (von Hippel, 2007), this approach is unlikely to produce bias, because we used one imputation model for imputing our dependent variables, depression and quality of life; furthermore, the dependent variables were highly correlated and were auxiliary variables for each other.

Another possible limitation is the use of five imputed datasets. Although five imputed datasets have been suggested to be sufficient it is now believed that using a larger number of imputation increases the precision of the results (Sterne et al., 2009). Another possible limitation of our study is that Stata does not allow random effects to be introduced in the imputation stage, which could be the reason for the

poor performance in estimating the between variance. Longitudinal data can be thought of as clustered or two-level data (Goldstein, 2003). It was suggested if a data set to be imputed is multilevel, then the imputation model should be multilevel too (Carpenter & Goldstein, 2004). There is no definitive recommendation in the literature on the best way to impute clustered data, but one of the strategies that has been suggested in Stata and that we have implemented is to impute all clusters simultaneously. The multiple imputations methods used in this work were performed using data in wide format, i.e. unstacked data with one observation row per subject, thus implicitly allowing an unrestricted covariance structure. Alternatives include the use of MICE package in R software (van Buuren & Groothuis-Oudshoorn, 2011), which allows for random effects in the imputation model. More recently, REALCOM-IMPUTE software, which performs multilevel multiple imputation, has been improved in order to allow missing values in covariates, including models where there are interactions, ordered and unordered categorical variables, or other functions of covariates such as polynomials (Goldstein, Carpenter, & Browne, 2014). Future work should consider comparisons of performance with this method.

It is possible that the approach adopted for the simulation study might have excluded variability from repeated sampling of the data values. To evaluate whether this would compromise the relative performance of the approaches used for missing data we have run a sensitivity analysis using a simulated data, which randomly sampled (with replacement) 1998 rows from our complete cases data set to create a new complete-cases sample. We have then applied our missing-value mechanism as before and stored the results. This process was repeated 100 times and the estimates were stored and summarized (results available on request). As expected, the ranking of each missing data approach did not change.

Although MICE and two-fold FCS procedures are useful when the specification of a joint multivariate distribution of all the variables with missing values is difficult, from a theoretical standpoint these techniques can be problematic because the sequence of regression models might not be consistent with a true joint distribution (Shafer & Graham, 2002). This means that the iterative algorithm might never converge because the joint distribution to which they might converge does not exist. Despite the lack of a satisfactory theory, MICE seems to work quite well in many applications. A number of simulation studies provide evidence that MICE and two-fold FCS generally yield estimates that are unbiased and that possess appropriate coverage, at least in the variety of cases investigated (Brand, van Buuren, Groothuis-Oudshoorn, & Gelsema, 2003; Raghunathan, Lepkowski, van

Hoewyk, & Solenberger, 2001; van Buuren, 2007; Nevalainen et al., 2009; Welch, Bartlett, & Petersen, 2014).

An anonymous reviewer suggested an extension to this study which employs a simulation study with a longer follow-up, and in which the following methods for dealing with missing data are included in the comparisons: multilevel imputation, imputation using long format, longitudinal weighting combined with complete case analysis, and longitudinal weights combined with multiple imputation. A further suggestion was an extension to this study which compares several missing at random mechanisms plausible in longitudinal studies and in cross-sectional studies, since the causes leading to attrition are different from those leading to item non-response. Future studies on multiple imputation in the context of panel data might usefully consider incorporating these suggestions.

It was shown when dealing with non-monotone missing data in longitudinal studies where a continuous outcome is involved, FIML and MI techniques all perform well. MI techniques compared to FIML might be more suitable for accommodating interaction terms. It was also shown MICE and two-fold FCS produced estimates that were more accurate and precise than those obtained from FIML and MVNI techniques, especially when dealing with non-continuous variables and interaction terms. The results of this study showed that MICE in general showed slightly better precision and accuracy better than two-fold FCS. More studies are needed to test the performance of two-fold FCS especially in the presence of many waves.

References

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67*(4), 1012-1028. doi: 10.1111/j.1741-3737.2005.00191.x

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37. doi: 10.1016/j.jsp.2009.10.001

Brand, J. P. L., van Buuren, S., Groothuis-Oudshoorn, K., & Gelsema, E. S. (2003). A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica, 57*(1), 36-45. doi: 10.1111/1467-9574.00219

Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior, 32*(1), 83-92. doi: 10.5993/ajhb.32.1.8

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*(24), 4279-4292. doi: 10.1002/sim.2673

Carpenter, J., & Goldstein, H. (2004). Multiple imputation using MLwiN. *Multilevel Modelling Newsletter, 16*(2), 9-18. Retrieved from http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/new16-2.pdf

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351. doi: 10.1037/1082-989x.6.4.330

Enders, C. K. (2001a). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement, 61*(5), 713-740. doi: 10.1177/0013164401615001

Enders, C. K. (2001b). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*(4), 352-370. doi: 10.1037/1082-989x.6.4.352

Ferro, M. A. (2014). Missing data in longitudinal studies: Cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. *Annals of Epidemiology, 24*(1), 75-77. doi: 10.1016/j.annepidem.2013.10.007

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, UK: Edward Arnold. doi: 10.1002/9780470973394

Goldstein, H., Carpenter, J., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 177*(2), 553-564. doi: 10.1111/rssa.12022

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician, 61*(1), 79-90. doi: 10.1198/000313007x172556

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association, 100*(469), 332-346. doi: 10.1198/016214504000001844

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Medical Research Methodology, 10*(7). doi: 10.1186/1471-2288-10-7

Minini, P., & Chavance, M. (2004). Sensitivity analysis of longitudinal binary data with non-monotone missing values. *Biostatistics, 5*(4), 531-544. doi: 10.1093/biostatistics/kxh006

Molenberghs, G., & Kenward. M. G. (2007). *Missing data in clinical studies*. Chichester, UK: John Wiley & Sons. doi: 10.1002/9780470510445

Nevalainen, J., Kenward, M. G., & Virtanen, S. M. (2009). Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification. *Statistics in Medicine, 28*(29), 3657-3669. doi: 10.1002/sim.3731

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6*(3), 328-362. doi: 10.1177/1094428103254673

Peyre, H., Leplege, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research, 20*(2), 287-300. doi: 10.1007/s11136-010-9740-3

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology, 27*(1), 85-96. Retrieved from http://www5.statcan.gc.ca/olc-cel/olc.action?ObjId=12-001-X20010015857&ObjType=47&lang=en

Royston, P., & White, I. (2011). Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of Statistical Software, 45*(4). doi: 10.18637/jss.v045.i04

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons. doi: 10.1002/9780470316696

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman and Hall/CRC.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. doi: 10.1037/1082-989x.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545-571. doi: 10.1207/s15327906mbr3304_5

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6*(4), 317-329. doi: 10.1037/1082-989x.6.4.317

Steptoe, A., Breeze, E., Banks, J., & Nazroo, J. (2013). Cohort profile: The English longitudinal study of ageing. *International Journal of Epidemiology, 42*(6), 1640-1648. doi: 10.1093/ije/dys168

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G.,… Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *The BMJ, 338*, b2393. doi: 10.1136/bmj.b2393

van Buuren, S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3), 219-242. doi: 10.1177/0962280206074463

van Buuren, S., & Groothuis-Oudshoorn, C. G. M., (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3). doi: 10.18637/jss.v045.i03

von Hippel, P. T. (2007). Regression with missing *Y*s: An improved strategy for analyzing multiply imputed data. *Sociological Methodology, 37*(1), 83-117. doi: 10.1111/j.1467-9531.2007.00180.x

Welch, C., Bartlett, J., & Petersen, I. (2014). Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. *The Stata Journal, 14*(2), 418-431. Retrieved from http://www.stata-journal.com/article.html?article=st0345

White, I., Royston, P., Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399. doi: 10.1002/sim.4067

Young, R., & Johnson, D. R. (2010). Imputing the missing Y's: Implications for survey producers and survey users. In *Proceedings of the AAPOR conference abstracts* (pp. 6242-6248).