Applications and Applied Mathematics: An International Journal (AAM)

Volume 13 | Issue 1 Article 6

6-2018

Multiple-Model Multiple Imputation for Longitudinal Count Data to Address Uncertainty in Missingness Mechanism

E. J. Farahani
Tarbiat Modares University

T. Baghfalaki
Tarbiat Modares University

Available at
http://pvamu.edu/aam
Appl. Appl. Math.
ISSN: 1932-9466

Applications and Applied Mathematics:
An International Journal
(AAM)

Vol. 13, Issue 1 (June 2018), pp. 84 - 96

Multiple-Model Multiple Imputation for Longitudinal Count Data to Address Uncertainty in Missingness Mechanism

E. Jalali Farahani and *T. Baghfalaki

Department of Statistics
Faculty of Mathematical Sciences
Tarbiat Modares University
Tehran, Iran
*E-mail: t.baghfalaki@modares.ac.ir

Received: August 14, 2016; Accepted: February 2, 2017

Abstract

In this paper, an approach to generate imputed values for count variables to incorporate missing data mechanism uncertainty is proposed. For multiple imputation, a distribution is considered in such a manner that it can reflect missing data mechanism uncertainty. For combining the parameter estimation of these imputed data sets the rules of nested multiple imputation are used. The performance of the multiple imputations is investigated using some simulation studies. Also, a real data set is analyzed using the proposed approach.

Keywords: Generalized estimating equations; Longitudinal study; Missingness; Multiple imputation methods; Sensitivity analysis

MSC 2010 No.: 62J12, 62P10

1. Introduction

A longitudinal study refers to an investigation where participant outcomes and possible treatments or exposures are collected at multiple follow-up times. Therefore, longitudinal studies generally yield multiple or repeated measurements on each subject over time. For example, HIV patients may be followed over time and some of their characteristics such as CD4 counts or viral load are collected to characterize the immune status and their disease burden, respectively. The repeated measurements for each subject are correlated within subjects and thus require special statistical techniques for valid analysis and inference. Longitudinal studies play a key role in epidemiology, clinical research and therapeutic evaluation. One of the major issues associated with the analysis of longitudinal data is the

existence of missing data or, more specifically, monotone missing data that arise when subjects dropout of the study.

Rubin (1976) distinguished between three important missing mechanisms. When missingness is unrelated to the data, missingness mechanism is termed missing completely at random (MCAR). When missingness depends on the observed data and when given the observed data, it does not depend on the unobserved data, the mechanism is missing at random (MAR). A mechanism where missingness depends on the unobserved data perhaps in addition to the observed data is termed missing not at random (MNAR). In the likelihood and Bayesian paradigm and when mild regularity conditions are satisfied, the MCAR and MAR mechanisms are ignorable, in the sense that inferences can proceed by analyzing the observed data only, without explicitly addressing a (parametric) form of the missing data mechanism. In this situation, MNAR mechanisms are nonignorable.

Ignoring the missing data mechanism may lead us to have overestimation or underestimation of parameters. Since a nonignorable missing data mechanism depends on unobserved data, there is little information available to correctly model the underlying process. A commonly used approach in such cases is to perform a sensitivity analysis drawing inferences based on a variety of assumptions regarding the missing data mechanism (Daniels and Hogan, 2008). There is a broad literature on sensitivity analyses for exploring unverifiable missing data assumptions (Ibrahim and Molenberghs, 2009). One approach begins with the specification of a full-data distribution, followed by examination of inferences across a range of values for one or more unidentified parameters (Daniels and Hogan, 2008; Molenberghs et al., 2001; Rubin, 1977; Scharfstein et al., 1999; Vansteelandt et al., 2006). When a decision is required, a drawback of sensitivity analysis is that it produces a range of answers rather than a single answer (Scharfstein et al., 1999). Several authors have proposed model-based methods for obtaining a final inference. This approach involves placing an informative prior distribution on the unidentified parameters that characterize assumptions about the missing data mechanism. Then, inferences are drawn that incorporate a range of assumptions regarding the missing data mechanism (Daniels and Hogan, 2008; Forster and Smith, 1998; Kaciroti et al., 2006; Rubin, 1977). An alternative approach for handling data with nonignorable missingness is the use of multiple imputations. Nested or two-stage imputation refers to multiple imputations conducted in a nested fashion. In the first stage, m imputations are generated. In the second stage, n imputations are generated for each completed data set in the first stage, resulting in a total of $M = mn$ multiple-imputed data sets. Then, using some combining rules which will be described in Section 2.4 the final inference is reported.

Siddique et al. (2013, 2014) described a new multiple imputation approach for estimating parameters and their associated confidence intervals in the presence of nonignorable nonresponse for continuous and binary variables. Their goal was to develop a multiple imputation framework analogous to model-based methods such as those of Rubin (1977), Forster and Smith (1998) and Daniels and Hogan (2008) that incorporate a range of ignorability assumptions into one inference. In this paper, we develop their method for count data by using multiple imputation models and combining rules. In this method there is a parameter that is unrecognizable. We try to specify a new algorithm for approximating an estimate of this parameter according to the observed and imputed values under missing at random mechanism.

This paper is organized as follows: nested multiple imputations for analyzing count data are described in the next section. This section contains three subsections with each subsection to be a part of the nested multiple imputation. In section 3, some simulation studies are performed

for investigating the performance of the proposed approach and in section 4 the missing values of a real data set using the proposed approach are imputed and then a real data set is analyzed. The last section includes some conclusions.

2. Nested Multiple Imputation for Analyzing Count Data

The approach proceeds in four steps as follows (Siddique et al., 2013):

1. Specification of a distribution of imputation model. In this step, after specification of the model, M model is drawn from this distribution of model.
2. Conducting nested multiple imputations which lead to N imputation for each model. In the end of this step $M \times N$ complete data sets are obtained.
3. Estimating parameters for each complete data set.
4. The use of nested multiple imputation rules for combining parameter estimation and standard errors of them. This step yields the final results for inference.

In what follows, these steps are discussed.

2.1. Specification of a distribution of imputation model

The first step of this approach is to identify the distribution for imputation. In fact, a good choice for this distribution is based on subjective information about association between missing values and observed data. The best information about this association can be gathered by experts and the persons who collected the data.

For continuous and binary data sets Siddiqe et al. (2013, 2014) proposed some approaches based on the ideas of Rubin (1987) for multiple imputations, assuming non-ignorability. Based on Rubin (1987, p. 203), there is a simple transformation for generating non-ignorable missing values from ignorable imputed values for continuous variable as follows:

$$(\text{non} - \text{ignorable imputed } Y_i) = \text{k} \times (\text{ignorable imputed } Y_i), \quad (1)$$

where Y_i is a continuous variable for the i^{th} individual, $i = 1,2,\dots,n$, and k is a constant multiplier. As an example from Rubin (1987) consider $k = 1.2$, this value shows that non-ignorable imputed values are 20% larger than those of ignorable imputed values or observed values. Based on this idea Siddique et al. (2013) proposed considering some distributions for k and drawing some values from this distribution to address missing data mechanism uncertainty. The proposed distribution of k is dependent on the imputer's belief about association of non-ignorable and ignorable missing values. For example, if the imputer believes that missing values tend to be larger than observed values, a proposal distribution for k might be $U(1,3)$ or $N(1.5,1)$. It is clear that this approach is appropriate for use, if in analyzing continuous variable and for other type of data, Equation (1) may generate implausible values. For binary data sets Siddique et al. (2014) proposed the following relationship for generating non-ignorable imputed values using ignorable values

$$\frac{\hat{\pi}_{non-ignor}/(1-\hat{\pi}_{non-ignor})}{\hat{\pi}_{ignor}/(1-\hat{\pi}_{ignor})} = k,$$

where $\hat{\pi}_{\text{non-ignor}}$ and $\hat{\pi}_{\text{ignor}}$ are the probability of the event under non-ignorability and the probability of the event under ignorability, respectively. Also, k is a constant multiplier which

shows the odds of the event for subjects with non-ignorable missing data as compared with ignorable missing data.

None of these approaches can be used for imputation of non-ignorable count missing values. We, instead, propose another approach. Let $Y_i \sim Pois(\lambda)$ and let $\lambda_{\text{non-ignor}}$ and λ_{ignor} be the mean parameter of non-ignorable missing values and ignorable missing values, respectively. Also, let $\log\ (\lambda_{\text{non-ignor}}) = \mu_{\text{non-ignor}}$ and $\log\ (\lambda_{\text{ignor}}) = \mu_{\text{ignor}}$. Then,

$$\lambda_{\text{non-ignor}} = k^* \times \lambda_{\text{ignor}}, \tag{2}$$

where k^* is a constant multiplier. Taking the logarithm of both sides, we obtain

$$\mu_{\text{non-ignor}} = k + \mu_{\text{ignor}}, \tag{3}$$

where $\log\ (k^*) = k$.

Using different values of k, which is generated by its assumed distribution, you would generate data from a model with λ values equal to λ values of an ignorable model multiplied by the value of the generated k.

2.2. Nested multiple imputation

After specification of the distribution of the models in the previous steps, imputation proceeds in two stages. First, M models are drawn from the distribution of the model. Then, N multiple imputations for each missing value are generated for each of the M models. Therefore, there are $M \times N$ complete data sets (Harel, 2007; Shen, 2000).

Let $Y = (Y^{obs}, Y^{mis})$ be a partition of the responses. In the first step, the imputation model ψ is drawn from its predictive distribution $\psi_m \sim p(\psi),\ m = 1,2,\ldots,M$. In the next stage for each model ψ_m, N independent imputations conditional on ψ_m, that is, $Y_{m,n}^{mis} \sim p(Y^{mis}|\psi_m)$, $n = 1,2,\ldots,N$ are drawn.

This kind of imputation is nested multiple imputation because the $M \times N$ observations are not independently drawn from the same posterior distribution. Therefore, the nested multiple imputation rules have to be used to take into account the variability due to the multiple models.

2.3. Estimating parameters for each complete data set using Generalized estimating equations

The method of Generalized Estimating Equations (GEE, Zeger et al., 1988) is a powerful approach for analyzing longitudinal data especially for longitudinal count, binary and ordinal responses. The regression coefficients and variance components in this method are estimated by the two first moments of asymptotic distribution of population and this method does not require the marginal distribution or likelihood function. Therefore, it is a widely used approach in medical and clinical data analysis. In this perspective, association among repeated measurements is considered by different structures of correlation matrix. A correct choice of the specification of the structure of correlation matrix is an approach for improving efficiency of regression coefficients.

Let $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ be a T-dimensional vector of the response variable and X_i is a $T \times p$-dimensional matrix of the explanatory variables for the i^{th} individual, $i = 1,2, \ldots, n$. In modeling count data using GEE, we assume that $E(Y_{ij}) = \lambda_{ij}$, $log(\lambda_{ij}) = X_{ij}\beta = \mu_{ij}$ and $var(Y_{ij}) = \phi\mu_{ij}$. In this structure X_{ij} is a p-dimensional vector of explanatory variable, β is a p-dimensional vector of regression coefficients and the logarithmic link function is used for modeling. Let $R_i(\alpha)$ be a $T \times T$ correlation matrix. The covariance matrix is given by $V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$ where $A_i = diag(V(\mu_{ij}))$ is a diagonal matrix with components $V(\mu_{ij})$. Estimates of the method of GEE for β ($\hat{\beta}$) in a marginal equation is given by solving the following equation (Liang and Zeger, 1986):

$$\sum_{i=1}^{I} D_i' \, V_i^{-1}(Y_i - \mu_i(\beta)) = 0, \tag{4}$$

where $D_i = \frac{\partial \mu_i}{\partial \beta}$. Note that, when the specification of the model is correct, $\hat{\beta}$ is a consistent estimate for β. Also, $n^{1/2}(\hat{\beta} - \beta)$ is asymptotically multivariate normal distribution as $n \longrightarrow \infty$ with mean vector 0 and covariance matrix

$$\left(\sum_{i=1}^{n} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta'}\right)^{-1} \left(\sum_{i=1}^{n} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \text{var}\ (Y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta'}\right) \left(\sum_{i=1}^{n} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta'}\right)^{-1}.$$

The parameter estimations and standard errors in using GEE approach for the resulting $M \times N$ complete data of our multiple models multiple imputation approach are computed in this stage.

2.4. Combining rules for final inference

Let β be the regression coefficient and let $\hat{\beta}_{m,n}$, $m = 1,2, \ldots, M$; $n = 1,2, \ldots, N$, be the estimated values from the resulting $M \times N$ complete data. Based on the large sample statement we have $\hat{\beta}_{m,n} - \beta \sim N(0, \Sigma_{m,n})$, $m = 1,2, \ldots, M$; $n = 1,2, \ldots, N$. Note that the subscript m, n represents the n^{th} imputed data set under m^{th} model. In describing the following rules, we use notation that follows closely to that of Shen (2000).
Let $\bar{\hat{\beta}}$ be the overall average for $M \times N$ estimation of β. It is given by

$$\bar{\hat{\beta}} = \frac{1}{NM} \sum_{m=1}^{M} \sum_{n=1}^{N} \hat{\beta}_{m,n},$$

Also, let the average of βs in the m^{th} model is given by:

$$\hat{\beta}^m = \frac{1}{N} \sum_{n=1}^{N} \hat{\beta}_{m,n}, \; m = 1,2, \ldots, M.$$

Three sources of variation contribute to the uncertainty in β: between model variance, within model variance, and the overall average of the variance estimates. They are given by

- between model variance:

$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}^m - \bar{\hat{\beta}})^2,$$

- within model variance:

$$W = \frac{1}{M(N-1)}\sum_{m=1}^{M}\sum_{n=1}^{N}(\hat{\beta}_{m,n} - \hat{\beta}^m)^2,$$

- overall average of the variance estimates:

$$\bar{\Sigma} = \frac{1}{NM}\sum_{m=1}^{M}\sum_{n=1}^{N}\Sigma_{m,n}.$$

The total variance is given by:

$$T = \bar{\Sigma} + (1+\frac{1}{M})B + (1-\frac{1}{N})W.$$

Note that, the interval estimation and other asymptotic evaluation about β are based on *t* distribution with ν degrees of freedom such that

$$T^{-1/2}(\bar{\hat{\beta}} - \beta) \sim t_\nu,$$

where

$$\nu^{-1} = \frac{1}{M-1}\left(\frac{(1+1/M)B}{T}\right)^2 + \frac{1}{M(N-1)}\left(\frac{(1-1/N)W}{T}\right)^2.$$

2.5. Specification of distribution of multiplier *k*

For specification of imputation models, one needs to determine distribution of multiplier k. In this Section, we describe an empirical approach for specification of this distribution. The parameter k is a sensitivity and non-identifiable parameter. Therefore, the specification of its distribution should be performed by researcher's subjective belief and available information.

When a normal distribution is considered for the distribution of k, we can choose lower and upper limits of k and then compute the mean and standard error as follows:

$$\mu_k = \frac{k_{lower} + k_{upper}}{2},\ \sigma_k = \frac{k_{upper} - k_{lower}}{4}.$$

Another empirical approach for approximating the distribution of k is the use of the following algorithm which is based on the information extracted from observed and the imputed values under missing at random mechanism, and it is useful for monotone missingness.

Let d_i be the dropout location of subject i. Then, $d_i \in \{2,\ldots,J+1\}$. Also, let y_{obs} be

observed data and $y_{imp-ignor}$ be the imputed values under missing at random mechanism. For specifying mean of multiplier k at time j, μ_{k_j}, consider subjects with $d_i \neq J+1$ (that is subjects with complete data are not considered).

Let $j = 2$, for finding μ_{k_2}, let $\bar{y}_{obs,2}$ be the mean of the observed responses for subjects with $d_i > 2$ and $\bar{y}_{imp-ignor,2}$ be the mean of imputed data with missing at random mechanism for subjects with $d_i = 2$. Therefore, an empirical estimate for μ_{k_2} is given by

$$\mu_{k_2} = \frac{\bar{y}_{obs,2}}{\bar{y}_{imp-ignor,2}}.$$

Generally, for time j the mean of the multiplier k is given as follows:

$$\mu_{k_j} = \frac{\bar{y}_{obs,j}}{\bar{y}_{imp-ignor,J}},\ j = 2,\ldots,J-1 \tag{5}$$

In the last time, all of the subjects with $d_i \neq J+1$ are dropout and mean of observed response for these subjects are not available. We compute multiplier k for the last time by:

$$\mu_{k_J} = \frac{\bar{y}_{obs,(J-1)} + (\bar{y}_{obs,(J-1)} - \bar{y}_{obs,(J-2)})}{\bar{y}_{imp-ignor,J}}.$$

3. Simulation Study

In this section, the performance of the proposed approach is investigated using some simulation studies. At first we generated a longitudinal count data with non-ignorable missing values. We generate a data set with sample size $I = 500$ and with $J = 5$ repeated measurements. Let $Y_i = (Y_{i1},\ldots,Y_{iJ})$ and $Y_{ij} \sim Pois(\lambda_{ij})$, $i = 1,2,\ldots,I$, $j = 1,\ldots,J$, where

$$\begin{aligned} log(\lambda_{ij}) = \beta_0 + \beta_1 Time_j + \beta_2 Trt_i + \beta_3(Trt_i \times Time_j) \\ + \beta_4(Drop_i \times Time_j) + b_i, \end{aligned} \tag{6}$$

where $Time_j = 0,1,2,\ldots,4$, Trt_i is equal to one for treatment group and zero for control group such that each group has 250 individuals and $Drop_i$ is an indicator variable for missingness $\beta_0 = 4$, $\beta_1 = -0.4$, $\beta_2 = 0.5$, $\beta_3 = -0.4$, and $\beta_4 = -0.8$. The random effects b_i has a normal distribution with zero mean and variance $\sigma^2 = 0.25$. For generating nonignorable missing values on y_{ij} at time points 1, 2, 3 and 4, subjects who have $Drop = 1$ are dropped out with probabilities $0.25, 0.50, 0.75, 1$, respectively.

For imputation of missing values according to the Equation (1) we first generated 200 imputations of each missing value by using predictive mean matching (PMM) method (Little and Rubin, 2002) which assumes the missing data are MAR. Therefore $\hat{\lambda}_{ignor}$ can be computed. Using the methods described in previous sections, the imputed values by PMM method (which assume an ignorable missingness) is transformed to imputed values under nonignorable misssingness.

Table 1. Result of imputation under two different assumptions about missing data mechanism

		MNAR-mechanism data generation		MAR-mechanism data generation	
Rate of missingness	Criterion	MAR	MNAR	MAR	MNAR
1/4	Σei2	1537.07	858.87	190.42	230.66
	percent of bias	15.7	4.1	0.1	3.6
1/2	Σei2	1752.36	926.15	162.40	172.45
	percent of bias	16.2	0.8	0.8	3.2
3/4	Σei2	2271.02	1016.11	209.83	243.16
	percent of bias	16.8	1.2	0.3	3.3
2/3	Σei2	1431.7	814.73	197.46	232.97
	percent of bias	14.5	3.7	0.1	3.4

Table 2. The results of simulation study of multiple imputation of longitudinal count data using multiple models.

Ignor assump.	Uncertainty	Model	Bias	RMSE	Σei2	Width of CI	$\widehat{\gamma}$
MAR	None	N(0,0)	0.087	0.121	4456.85	0.08	0.008
	Mild	N(0,0.1)	0.090	0.123	4284.46	0.083	0.046
	Moderate	N(0,0.2)	0.093	0.125	4211.09	0.086	0.143
	Ample	N(0,0.3)	0.096	0.127	4157.77	0.090	0.229
MNAR	None	N(μ_k,0)	0.062	0.081	1493.41	0.097	0.200
	Mild	N(μ_k,0.1)	0.062	0.081	1550.42	0.110	0.397
	Moderate	N(μ_k,0.2)	0.062	0.082	1618.34	0.124	0.527
	Ample	N(μ_k,0.3)	0.061	0.082	1702.98	0.143	0.637

Specifically, we simulated 100 values of k from normal distribution with mean based on the above explained method. We used $M = 100$ imputation models and $N = 3$ imputations within each model so that the degrees of freedom for the within-model variance is $M(N-1)$ and the degrees of freedom for the between-model variance is $M-1$. This allows us to estimate within and between-model variance with equal precision, which is necessary for stable measurements of the rates of missing information (Harel, 2007).

We explored the effect of imputing under two different ignorability assumptions which we refer to as MAR and MNAR. In addition to generating imputations using the above ignorability assumptions, we also generated imputations based on four different assumptions regarding how certain we were about the correctness of our models. When there is no mechanism uncertainty, all imputations are generated from the same model. When there is mechanism uncertainty, then multiple models are used. All models are centered on one of the ignorability assumptions. The four different uncertainty assumptions used to generate multiple models were as follows: no uncertainty, mild uncertainty, moderate uncertainty and ample uncertainty.

Table 1 shows the results of imputation under MAR and MNAR mechanism with four different rates of missingness for generating data under missing at random and missing not at random mechanism. For comparison of the results, the values of sum of square errors (SSE) of imputed values and percent of bias for estimating regression coefficient for treatment parameter in Equation (6), are reported. This table shows that the values of MSEs and percent of bias under MNAR are smaller to those under MAR mechanism.

We then analyzed the 200 imputed data sets using the described model of Equation (6) after removing dropout as an explanatory variable and estimating the regression coefficient of the treatment group for imputed values.

The results of this simulation are summarized in Table 2. We evaluated the bias, Σe_i^2 and root of MSE (RMSE) of the treatment slope as well as width of its nominal 95% interval estimate. Also, missing data information is calculated. The missing data information for nested multiple imputations will be obtained according to the amount of missing information due to the uncertainty in the model and missing data. Let $\bar{Q}$ be the mean of $M \times N$ point estimates of parameter, $\bar{U}$ be the mean of estimated variances, B be between model variance and W be within model variance. Then, an estimate of missing data information, γ, will be equal to

$$\hat{\gamma} = \frac{B+(1-\frac{1}{N})W}{\bar{U}+B+(1-\frac{1}{N})W}.$$

Table 2 presents the results of our imputations under the 8 different ignorability/uncertainty scenarios using PMM imputation and the methods described for the slope of the treatment group. The first row shows the results of assuming MAR with no mechanism uncertainty, where the results are highly biased. Also, in this status missing data information, γ, is too small and Σe_i^2 is very large. The results show that with increasing mechanism uncertainty both the percentage of bias and RMSE approximately are the same as those under no uncertainty, but coverage and missing data information are increased with increasing uncertainty in the imputation models. Also, the results show that the values of percent of bias and RMSE under MNAR are smaller than those under MAR.

Table 3. Results of imputation missing values in AIDS data

Ignor assump.	Uncertainty	Model	Estimate	SE	Width of CI	p-value	$\widehat{\gamma}$
MAR	None	N(0,0)	12.10	13.22	27.75	0.18	0.02
	Mild	N(0,0.1)	12.10	17.18	58.13	0.33	0.09
	Moderate	N(0,0.2)	12.12	19.03	70.01	0.41	0.12
	Ample	N(0,0.3)	12.13	30.94	106.11	0.52	0.18
MNAR	None	N(μ_k,0)	8.10	11.41	30.71	0.23	0.01
	Mild	N(μ_k,0.1)	8.12	15.29	41.99	0.29	0.07
	Moderate	N(μ_k,0.2)	8.11	19.04	32.40	0.36	0.13
	Ample	N(μ_k,0.3)	8.09	24.54	99.12	0.57	0.20

4. Application

In this section, descriptive and inferential (modeling approach) methods are used for analyzing data sets of a longitudinal HIV study. The study contains 467 HIV infected patients who had failed or were intolerant of zidovudine (AZT) therapy. The data had been analyzed before by Ganjali and Baghfalaki (2014). The aim of their study was to compare the efficacy and safety of two alternative antiretroviral drugs, namely didanosine (ddI) and zalcitabine (ddC). Patients were randomly assigned to receive either ddI or ddC, and CD4 cell counts were recorded at study entry, where randomization took place, as well as 2, 6, 12, and 18 months thereafter. In order to impute missing values, we use PMM approach and two imputations are considered. These imputations are under missing at random assumption. For converting imputed values to some values under not missing at random, 100 values from distribution $N(\mu_k, \sigma_k^2)$ are

generated and replace k and $\hat{\lambda}_{ignor}$ in Equation (2) and create 2 imputations nested within 100 models; that is, 200 imputed data sets are generated. Like simulation study find μ_k by Equation (5). Again we consider four different amounts (0, 0.1, 0.2, 0.3) for σ_k^2. The location parameter of GEE equation considered as follow:

$$\mu_{ij} = \beta_0 + \beta_1 Time_j + \beta_2 Trt_i + \beta_3 Gender_i + \beta_4 \text{Prevoi}_i + \beta_5 \text{Stratum} + \beta_6 (Time_j \times Trt_i) \quad (7)$$

where $Time_j = 0, 2, 6, 12, 18$, $Gender_i$ is a gender indicator for the individual (0 = female, 1 = male) and the other three explanatory variables are: Trt (0 = ddC, 1 = ddI), Prevoi, previous opportunistic infection (1 = AIDS diagnosis, 0 = no AIDS diagnosis), and Stratum (1 = AZT failure, 0 = AZT intolerance).

For comparing the results under different missingness mechanism and uncertainty mechanism, this model is fitted and the regression coefficient of treatment group are estimated and reported in Table 3. The results show that with increasing uncertainty in imputation process, similar to that of simulation study, there is not much change in the parameter estimates but standard errors, width of confidence intervals, p-values and $\hat{\gamma}$ are increased.

Table 4. Results of estimation regression coefficients in AIDS data set under two different missingness mechanism

Ignor assump.	Uncertainty	Model	Parameters	Estimate	SE	Width of CI	p-value	$\hat{\gamma}$
MAR	None	N(0,0)	Intercept	58.19	6.91	21.37	0.03	0.29
			Time	-1.21	0.25	1.10	0.05	0.76
			Trt	12.10	13.22	27.75	0.18	0.02
			Gender	19.42	26.60	101.49	0.77	0.10
			Prevoi	-31.32	9.30	37.70	0.06	0.77
			Stratum	1.80	7.27	18.68	0.92	0.08
			Time×Trt	-3.14	1.16	2.26	0.03	0.44
MNAR	Moderate	N(μ_k,0.2)	Intercept	123.42	16.08	61.48	0.01	0.31
			Time	-1.11	0.19	0.86	0.04	0.83
			Trt	8.12	15.29	41.99	0.29	0.07
			Gender	-12.36	20.39	83.32	0.64	0.11
			Prevoi	-70.75	21.11	53.78	0.05	0.63
			Stratum	-1.14	6.45	16.35	0.84	0.09
			Time×Trt	-0.01	0.75	2.44	0.96	0.06

Table 4 lists the results of the estimated regression coefficients in Equation (7) for two models, one under MAR mechanism without uncertainty and the other under MNAR mechanism with mild uncertainty. The results show that there is considerable difference between the regression coefficients under the missingness assumptions. Also, Figure 1 shows longitudinal profiles for observed values of the response variables for each individual over time with black color and those with imputation values with red color. Panel (a) of this figure is under MAR mechanism without uncertainty and in panel (b) it is for under MNAR mechanism with mild uncertainty. The results show that considering MAR mechanism for imputation data in last time, a sharp decrease in the amount of imputation occurred that seems unreasonable but by changing ignorability mechanism to MNAR this problem is solved.

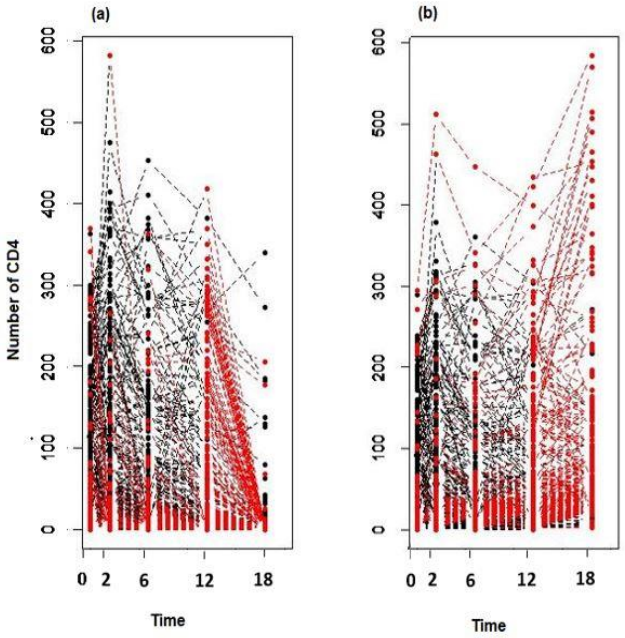

Figure 1. Profiles of CD4 measurements over time for observed values (black color) an imputed values (red color). Panel a: under missing at random assumption without uncertainty, panel b: under missing not at random assumption with mild uncertainty.

5. Conclusions and Discussion

In this paper we have developed the proposed approach of Siddiaqe et al. (2013, 2014) for generating multiple imputations for longitudinal count data with missing values. In this approach using multiple models and multiple imputations we have taken into account the uncertainty about the missing data mechanism in imputation process.
The full data distribution can be factored into an extrapolation model and an observed data model,

$$p(y, r|\omega) = p(y_{mis}|y_{obs}, r, \omega_{mis})p(y_{obs}, r|\omega_{obs})$$

where ω_{mis} and ω_{obs} denote parameters indexing the missing and observed data models, respectively. The observed data distribution is identified and can be estimated non-parametrically but the missing data distribution cannot be identified without modeling assumptions or constraints on the parameter space. To formalize this notion, we define a class of parameters for full-data models that can be used for sensitivity analysis or incorporation of informative prior information. Generally, they are not identifiable from observed data, but when their values are fixed, the remainder of the full-data model is identified, and we call them sensitivity parameter. Our use of the term sensitivity analysis refers to assessment of sensitivity of model-based inferences and to assumptions that cannot be verified or checked with data. Without assumptions such as a parametric model for the full-data response, or constraints such as MAR for the missing data mechanism, the observed data provide no information about the missing data distribution. The general strategy here is to work with the subset of sensitivity parameters (like multiplier k in this paper). The sensitivity parameters are then used to encode prior beliefs about the missing data mechanisms, either by fixing their values at some constant, examining inferences across a range of constants, or by assigning an appropriate prior distribution, Daniels and Hogan (2008).

As seen in both the simulation studies and the application, post-imputation inferences can be highly sensitive to the choice of the imputation model. When choosing a distribution for the multiplier k in $\lambda_{\text{non-ignor}} = k^* \times \lambda_{\text{ignor}}$ we described a method for the determination of distribution's parameter by $\mu_{k_j} = \frac{\bar{y}_{obs,j}}{\bar{y}_{imp-ignor,j}}$. As a future study, the observations may be classified by their covariates and the distribution's parameter for multiplier k may be defined in each category. For example subjects in the same treatment group or with the same gender have more similarities; therefore, μ_{k_j} in Equation (5) can be calculated for each category, separately. Also, for responses in exponential family one may use the same approach considering the canonical parameter of the distribution as a link. This would be an extension of all forms presented in this paper.

Some other approaches for generating multiple-model multiple imputations that can be incorporated into our framework include mixture model imputation (Rubin, 1987, van Buuren, Boshuizen and Knook, 1999), imputation based on a multivariate t-distribution with varying degrees of freedom (Liu, 1995) and pattern-mixture model imputation (Demirtas and Schafer, 2003, Thijs et al., 2002).

REFERENCES

Daniels, M. J. and Hogan, J. W. (2008). Missing Data in Longitudinal Studies. Monographs on Statistics and Applied Probability 109. Chapman & Hall/CRC, Boca Raton, FL.

Demirtas, H., and Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. Statistics in medicine, 22(16), 2553–2575.

Forster, J. J. and Smith, P. W. F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. Journal of the Royal Statistical Society: Series B, 60, 57-70.

Ganjali, M., and Baghfalaki, T. (2014). A Bayesian shared parameter model for analysing longitudinal skewed responses with nonignorable dropouts. International Journal of Statistics in Medical Research, 3(2).

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. Statistical Methodology, 4(1), 75-89.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. TEST, 18, 1-43.

Kaciroti, N. A., Raghunathan, T. E., Schork, M. A., Clark, N. M. and Gong, M. (2006). A Bayesian approach for clustered longitudinal ordinal outcome with nonignorable missing data: Evaluation of an asthma education program. Journal of the American Statistical Association, 101, 435-446.

Liang, K.Y. and Zeger, S.L., (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), pp.13-22.

Liu, C. (1995). Missing data imputation using the multivariate t distribution. Journal of multivariate analysis, 53(1), 139-158.

Molenberghs, G., Kenward, M. G. and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. Journal of the Royal Statistical Society: Series C, 50, 15-29.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63, 581-592.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in

sample surveys. Journal of the American Statistical Association, 72, 538-543.
Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley, New York.
Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94, 1096-1146.
Shen, Z. J. (2000). Nested multiple imputation. Ph.D. thesis, Dept. Statistics, Harvard Univ., Cambridge, MA.
Siddique J, Harel O, Crespi CM. (2013). Addressing missing data mechanism uncertainty using multiple model multiple imputation: application to a longitudinal clinical trial. Annals of Applied Statistics, 6, 1814–1837
Siddique J, Harel O, Crespi CM. (2014). Binary variable multiple-model multiple imputation to address missing data mechanism uncertainty using: application to a smoking cessation trial. Statistics in Medicine, 33, 3013–3028.
Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. Biostatistics, 3(2), 245-265.
Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicine, 18(6), 681-694.
Vansteelandt, S., Goetghebeur, E., Kenward, M. G. and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. Statistica Sinica, 16, 953-979.
Zeger S, Liang K, Albert P (1988) Models for longitudinal data: a generalized estimating equation approach. Biometrics, 44, 1049-1060