Open Journal of Statistics, 2014, 4, 933-944
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.411088

Comparison of Four Methods for Handing Missing Data in Longitudinal Data Analysis through a Simulation Study

Xiaoping Zhu

Biostatistics & Data Management, Regeneron Pharmaceuticals, Inc., Basking Ridge, USA
Email: Xiaoping.zhu@regeneron.com

Received 15 October 2014; revised 12 November 2014; accepted 22 November 2014

Abstract

Missing data can frequently occur in a longitudinal data analysis. In the literature, many methods have been proposed to handle such an issue. Complete case (CC), mean substitution (MS), last observation carried forward (LOCF), and multiple imputation (MI) are the four most frequently used methods in practice. In a real-world data analysis, the missing data can be MCAR, MAR, or MNAR depending on the reasons that lead to data missing. In this paper, simulations under various situations (including missing mechanisms, missing rates, and slope sizes) were conducted to evaluate the performance of the four methods considered using bias, RMSE, and 95% coverage probability as evaluation criteria. The results showed that LOCF has the largest bias and the poorest 95% coverage probability in most cases under both MAR and MCAR missing mechanisms. Hence, LOCF should not be used in a longitudinal data analysis. Under MCAR missing mechanism, CC and MI method are performed equally well. Under MAR missing mechanism, MI has the smallest bias, smallest RMSE, and best 95% coverage probability. Therefore, CC or MI method is the appropriate method to be used under MCAR while MI method is a more reliable and a better grounded statistical method to be used under MAR.

Keywords

MCAR, MAR, Complete Case, Mean Substitution, LOCF, Multiple Imputation

1. Introduction

The problem of missing observations can frequently occur in all types of clinical trials, especially when observations are measured repeatedly at each scheduled visit for the same subject in a longitudinal study. In longitu-

dinal studies, there are many possible causes leading to missing data including the duration of the study, the nature of the disease, the efficacy and adverse effects of the drug under study, accidents, patients' refusal to continue, moving, or other administrative reasons. Frequently, missingness can potentially lead to two serious problems in statistical practice: reducing the overall statistical power and having biases in the estimates. In statistical practice, missing data is a key problem that can never be avoided completely. Since the most traditional statistical methods are intentionally designed to handle complete data sets by default, therefore data analysts should pay special attention to incomplete data sets.

Little and Rubin [1] have classified missing data mechanisms into three different types based on the possible causes: 1) missing completely at random (MCAR) if the missingness is not related to any observed and unobserved factors (such as domestic relocation, suffering an accident, or unrelated illness); 2) missing at random (MAR) if the missingness is conditional on observed factors and is independent of the unobserved data (such as lack of efficacy); and 3) missing not at random (MNAR) if the missingness depends on unobserved quantities as well as some observed factors. The MNAR missing mechanism is usually used to describe patients who may drop out as a result of health deterioration related to the treatments that we do not have a chance to observe because of their dropout. Researchers have pointed out that the MAR assumption may be more plausible in practice than that of the MCAR [2]. In fact, by definition MCAR is only a special case of MAR. In other words, a MCAR missing mechanism is also a MAR one, but not every MAR is a MCAR. Actually, it is possible to formally test the MCAR assumption against its alternative hypothesis not MCAR [3] [4]. However, it is not possible to test MAR or MNAR without using additional (external) information. MNAR is particularly useful in assessing the sensitivity of the results that are not MAR [5] and it is highly recommended to be incorporated into the analysis.

In the literature, several alternative statistical approaches have been applied to the analysis of longitudinal data with missing values. These appropriate methods for analysis should be selected based on the data missing mechanism, since different statistical methods are valid only under certain situations (missing mechanisms) with specified missing rates. In other words, there is no unique best method available for all situations. However, it is difficult to test the missing mechanism in a longitudinal clinical study and there is also no clear rules regarding how much is qualified as too much missing data [6]. In general, the choice of a particular method for handling missing data depends largely on whether one is considering a more pragmatic or a more explanatory perspective. There is often the question of whether there are too many missing data. Sprint and Dupin-Sprint [7] pointed out that the tolerable amount of missing data is that would not conceal an effect in the opposite direction. In order to determine whether this level of missing data has been reached, one can perform what was called the "worst case" analysis.

Despite these difficulties, several researchers have considered and constructed simulation studies for the proof of strong consistency of imputation methods to check the efficiency of the imputation methods. For example, Myers [8] compared the results of two imputation methods (that is, the complete case method and the multiple imputation method) based on simulated data sets with a dropout rate ranging from 20% to 60%, and they concluded that MI method provided results that are more closely mimicked the complete data set.

Hening and Koonce [9] investigated five imputation methods (*i.e.*, mean substitution, median substitution, zero value, hot-deck, and MI) and a first-year-student retention data with more than 20% missing values is used. The results shown that multiple and hot-deck imputations perform poorly in an accuracy comparison test, but they can slightly increase the predication accuracy rate compared with other methods.

Ali, *et al*. [10] performed a survival analysis in which missing data were simulated under MCAR and MAR to compare four imputation methods—complete case analysis (CCA), means substitution (MS), and multiple imputation (MI)with the inclusion of the outcome (MI− and MI+). The simulation results suggested that in general MI+ is likely to be the best method. Patrician [11] pointed out that MI is the best approach and should be considered to handle missing data compared with CCA and MS by an empirical investigation of AIDS care longitudinal data outcomes.

Recently, Nakai, *et al*. [12] have shown that MI is the most effective imputation method in longitudinal data setting under MCAR via a simulation study. This indeed provides useful information about the performance of imputation methods under MCAR, but it is limited and restricted to clinical situations where MAR is more plausible. For example, Lavori, *et al*. [13] have pointed out that the MCAR assumption is often not plausible in

most clinical trial settings. The purpose of this paper is through a simulation approach to analytically evaluate the performance of four imputation methods for different missing mechanisms (MCAR and MAR) with various missing rates. For simplicity and also without loss of generality, a monotone pattern of missing data (meaning that once a patient has a missing response at an assessment visit, his or her data will be missing for all subsequent visits) is assumed. Under such assumptions, this paper primarily concentrates on the following four imputation methods: 1) complete case (CC); 2) mean substitution (MS); 3) last observation carried forward (LOCF); and 4) multiple imputation (MI). To compare the performance of these methods, bias, RMSE, and 95% coverage probability (CP) of the estimated parameters are used as evaluation criteria.

This paper is organized as follows. Section 2 reviews methods of missing data analysis. The simulation procedures (with available covariates) under MCAR and MAR settings are described in Section 3. In Section 4, the simulation results are used to evaluate the performance of those four imputation methods considered. Finally, discussion and concluding remarks are provided in Section 5.

2. Approaches to Handling Missing Data

There are so many techniques in handling missing data discussed in the literature. Especially, many methods have been proposed and developed to handle missing data in longitudinal clinical trials. However, there are few methods that are actually used in real trials with missing data. The purpose of this paper is to study four most frequently used methods for dealing with missing data and they will be described as follows.

2.1. Complete Case (CC) Analysis

This method deletes all cases with missing data and then performs statistical analyses on the remaining complete data set (which has a smaller sample size). Since all cases containing missing data have been removed, there is no missing data problem to handle. Therefore, all statistical methods can be used to analyze the smaller data set. Obviously, one major advantage of this method is its ease of use. In fact, virtually all statistical programs incorporate this method as a default method because it accommodates any type of statistical analysis [14]. The method may be preferred under the situation in which the sample size is large, the proportion of missing data is small, and the missing data mechanism is MCAR [15]. For MCAR missing data, the method will yield unbiased parameter estimates and larger standard errors due to the smaller sample size. However, even when data are MCAR, loss of data will result in loss of precision (larger standard errors), particularly in multivariate data analyses.

In general, the major disadvantage of the method is that it could possibly lead to losing statistical power due to the reduction of the sample size. Also, complete case techniques erode efficiency such that the variation (*i.e.*, the standard error) around the true estimate is too large [16]. In addition, if data are not MCAR, bias can be a serious issue [17]-[19].

2.2. Mean Substitution (MS)

The method of mean substitution imputes the missing values using the mean of the available observed values. This method has the potential of introducing biases as well as underestimating variability [20]-[22]. This method has the advantage of being able to maintain the original sample size while it also allows one to use the complete-data methods for data analysis [23]. However, due to the reduced variability, the estimated parameters are less precise. Decreased variances are problematic because the resulting estimates are too close to the mean [24].

2.3. Last Observation Carried Forward (LOCF)

The simplest imputation approach is the LOCF method that replaces every missing value with its corresponding last observed value. LOCF method is often used in longitudinal studies of continuous outcomes under MCAR. Conceptually, this method assumes that the outcome would not change after the last observed value. Therefore, there is no time effect since the last observed data. In fact, LOCF has been a popular method that is frequently used in handling missing data problems because it is easy to understand and can be implemented easily as well. Also, unlike the CC method, the sample size does not change. For example, in a clinical trial (see the data below), patient 3 dropped out from the study after baseline. Patient 6 dropped out after the first month follow up.

Subject	Group	Baseline	Follow-Up Week		
			1st Month	3rd Month	6th Month
1	Placebo	296	175	187	192
2	Placebo	376	329	236	76
3	Placebo	150	?(150*)	?(150*)	?(150*)
4	Active	282	186	225	134
5	Active	317	31	85	120
6	Active	362	104	?(104*)	?(104*)

?: missing value; *: value imputed by LOCF.

If there are more patients dropped out in the placebo group due to the lack of efficacy, then this method might give a biased conclusion about the effect of the treatment group. In general, the measurements are unlikely to remain unchanged for either placebo or treatment group. In our example, the measurement of patient 3 from the control group will increase while that of patient 6 from the active group will decrease. This implies that there is no improvement in the active group and hence no difference between these two groups.

Rigorously speaking, LOCF is not an analytic approach, but it is a method that is very easy to impute missing values. Analytic proofs [25] [26] and studies in simulated data [26]-[29] have been clearly shown that LOCF can bias results and lead to either overestimation or underestimation of the parameter estimates.

2.4. Multiple Imputation (MI)

Multiple imputation was first proposed by Rubin [23] [30] to analyze incomplete data under the MAR missing mechanism. Allison [14] and Schafer [31] [32] further explained Rubin's MI. MI is a predictive approach to handling missing data in multivariate analysis and blends both classical and Bayesian statistical techniques. The idea of MI is to impute a missing value multiple times and hence generates multiple (m) data sets. Missing data theorists have claimed that good inferences can be made for only 3 - 5 imputed data sets [33]. Others have suggested that the number of imputed data sets should approximate the percentage of subjects with some missing data [34]. In this paper, 100 imputed data sets are used because there is a 50% missing rate in Case 3 of the simulation study. Then, these imputed data sets are analyzed by standard procedures that are commonly used in analyzing complete data sets. Finally, the results of analyses are combined [35] [36]. A few assumptions and constraints of MI are: 1) missing data mechanism should be MAR; 2) the imputation model must match the analysis model [14]; and 3) the algorithm used to generate imputed values must accommodate/include the variables associated with the missingness of the data as well as other related variables. Allison [37] illustrated that good imputation methods use all information related to missing cases. Two major advantages of MI are allowing the use of complete-data methods for data analysis and incorporating random errors in the imputation process. MI can accommodate any model with any data and does not require specialized software. In addition, MI increases efficiency of the estimates through minimizing the standard errors [23]. However, Rubin [23] pointed out that the three disadvantages of MI are more effort to create the multiple imputations, more time to run the analyses, and more computer storage space for the imputation-created data sets.

3. Simulations

We conducted a simulation study on different scenarios. In general, generating each dataset is based on the setting described in Section 3.1 by the following assuming: 1) the measurement at the first time point (t = 1) from the original data is completely observed; 2) data are MCAR and MAR missing mechanisms; 3) the missing pattern is monotone. To begin the simulated process, the first step (1-step) generates the five-time points of measurements of each subject by a random number from a multivariate normal distribution with AR(1) correlation structure and repeats the step 100 times for 100 subjects, given the observed values, $Y_{it(obs)}$; $i = 1, 2, \cdots, 100$; and $j = 1, 2, \cdots, 5$. The second step (2-step) is to generate the MCAR and MAR data, $Y_{it(miss)}$, as described in Section 3.3 by the missing rate (%) at each time point measurement in **Table 1**. The third step is to test the MCAR and MAR condition using Little's MCAR test to check whether the produced datasets are MCAR and MAR or not.

Table 1. Detailed design of the simulation.

	n	β_1	β_0	σ^2	ρ	Missing Rate at Time t				
						1	2	3	4	5
Case 1	100	0.1 2 10	10	1	0.7	0%	5%	5%	5%	5%
Case 2	100	0.1 2 10	10	1	0.7	0%	5%	20%	15%	20%
Case 3	100	0.1 2 10	10	1	0.7	0%	10%	20%	30%	50%

Finally, each of the predefined 9 situations was repeated 1000 times by using SAS procedures. There are 1000 data sets each containing 100 subjects with 5 time points per subject to be analyzed. Next, the standard mixed model procedures were performed on these simulated data sets. Finally, the regression coefficients and their standard errors were obtained. Then, the performance of four selected methods (that is, complete case analysis, mean substitution, last observation carried forward, and multiple imputation) was compared based on biases, root mean square errors, and 95% coverage probabilities. In the simulation, we considered the missing data with MCAR and MAR missing mechanisms. In addition, without loss of generality, the missing pattern was assumed to be monotone. The missing rate varied from 5% to 50%.

3.1. Background of the Simulation

In the simulation, we generated the longitudinal data Y_{it} $(i=1,2,\cdots,100;\ j=1,2,\cdots,5)$ for the i^{th} subject at the t^{th} visit according to a multivariate normal distribution model, $E(Y_{it})=\beta_0+\beta_1 t$ where β_0 is the intercept and β_1 is the slope. The variance at each occasion is assumed to be constant over time, while the correlation coefficient between Y_{is} and Y_{it} is assumed to be a positive correlation coefficient ρ of a first-order autoregressive model (*i.e.*, $AR(1)$).

More precisely, a data set X with n rows and p columns is drawn from a multivariate normal distribution with a zero mean vector and a variance-covariance matrix $\boldsymbol{\Sigma}$ given as follows

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_{1,1} & \cdots & \cdots & \cdots & \sigma_{1,5}\\ \vdots & \ddots & \cdots & \cdots & \cdots\\ \vdots & \cdots & \ddots & \cdots & \cdots\\ \vdots & \cdots & \cdots & \ddots & \cdots\\ \sigma_{5,1} & \cdots & \cdots & \cdots & \sigma_{5,5}\end{bmatrix}$$

The correlation structure of an $AR(1)$ model can be described as

$$R_{AR(1)}=\begin{bmatrix}1 & \rho & \rho^2 & \cdots & \rho^4\\ \vdots & 1 & \vdots & \vdots & \vdots\\ \vdots & \vdots & 1 & \vdots & \vdots\\ \vdots & \vdots & \vdots & 1 & \vdots\\ \rho^4 & \cdots & \cdots & \cdots & 1\end{bmatrix}$$

where

$$\rho_{i,t}=\frac{\sigma_{i,t}}{\sqrt{\sigma_{i,i}}\sqrt{\sigma_{t,t}}},\ i=1,2,\cdots,100;\ t=1,2,\cdots,5$$

In this study, the correlation coefficient ρ is taken to be 0.7 to simulate the strong relationships among variables. The total number of cases (or subjects) is 100 and there are 5 measurements at 5 time points for each subject.

3.2. Design of the Simulation

Table 1 listed the values of parameters used in the simulation study. The value of the intercept was fixed at 10 while the value of the slope was 0.1, 2, or 10. The variance was chosen to be 1. And, the correlation coefficient was set to 0.7, where $\rho = 0.7$ is a typical value used to represent moderate to high correlation. Nakai (2014) used both $\rho = 0.1$ and $\rho = 0.7$ in his study. However, he showed the results are not significantly different for both values of correlation coefficient. Three combinations of missing rates under both MCAR and MAR missing mechanisms were considered. (See **Table 1** for details).

Therefore, in the simulation 1000 samples were generated covering 18 different situations: 3 (missing rates) × 3 (values of the slope) × 2 (missing mechanisms).

3.3. Missing Data Generation

The data were generated with situations described in Sections 3.1 and 3.2 and the measurements were drawn from a multivariate normal distribution with *AR*(1) correlation structure [12]. The data generating process was repeated 1000 times for each of 18 different situations. PROC IML of the SAS System (version 9.3) was used to generate these data sets. Hereafter, these simulated data sets (without any missing observation) are referred to as the "Original" data sets.

After the original data sets were created, the measurements at different time points for different subjects were set to missing, according to the MCAR or MAR missing mechanism. However, the measurement at the first time point (that is, the baseline value) of each subject was assumed always observed. In the MCAR setting, missing data were generated randomly at visits 2 through 5 based on the missing probabilities listed in **Table 1**. Therefore, the missing probabilities do not depend on either observed or unobserved data. Furthermore, Little's MCAR test [3] was performed to make sure the missing mechanism is indeed MCAR otherwise that data set was discarded and another data set was generated anew.

In the MAR setting, the probability of missing at visit 2 was set in proportional to the baseline values based on a logistic probability distribution model. In the same way, the missing data at visits 3 through 5 were set based on the probabilities given in **Table 1**. In this way, the missing probabilities will depend only on observed data. Hence, the missing mechanism is under MAR mechanism by MAR definition. It was quite time-consuming in generating and analyzing these data sets. Actually, it would take almost a day to do such a job for a single simulation setting.

3.4. Measures of Performance for Imputation Methods

Bias, root mean squared error (RMSE), and coverage probability are used as criteria to assess the performance of the four imputation methods. The SAS System (version 9.3) is used to perform all the statistical analyses as well as to produce the required results. Also, we set covariance structure to "Unstructured" simply to explore the accuracy of imputation within "PROC MIXED" procedure. These criteria will be described in detail in the following subsections.

3.4.1. Bias

Bias is defined as the difference between the average value of estimated parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$) and the true parameters (β_0 and β_1) obtained from the corresponding original data set.

3.4.2. RMSE

The mean squared error (MSE) is defined as the average squared difference between the estimated parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$) and the corresponding true parameters (β_0 and β_1) obtained from the original data set. MSE is equal to the sum of variance and the squared bias of the estimated parameters. RMSE is defined as the square root of the MSE. The RMSE is a useful measure of overall precision or accuracy and can be used to evaluate the performance of each imputation method. In general, the more effective method would have a lower RMSE [38].

3.4.3. Coverage Probability

The coverage probability (CP) is defined as the proportion of the simulated data sets, among the 1000 simulated data sets, that yield the 95% confidence intervals containing the true parameter values based on the original data sets. Therefore, an appropriate method should have a coverage probability around 95%.

4. Simulation Results

The simulation results are summarized in **Tables 2-7**. In these table, CC, MS, LOCF, MI, MCAR, and MAR stand for complete case, mean substitution, last observation carried forward, multiple imputation, missing completely at random, and missing at random, respectively. Case 1, Case 2, and Case 3 represent the low, moderate, and high missing rate setting as given in **Table 1**. In the following tables, bold numbers are used to highlight the best method in that particular case.

4.1. Simulation Results for MCAR Missing Data

Tables 2-4 show the simulation results for MCAR (missing complete at random) data. The cases of low, moderate, and high missing rates are illustrated in **Table 2**, **Table 3**, and **Table 4**, respectively. In the low missing rate case (**Table 2**), except the LOCF method, the other three methods yield very small biases and RMSEs for both Intercept and Slope. In the case that the slope is 10, the 95% CPs are very poor for the LOCF method (100% and 0% for Intercept and Slope, respectively). In the moderate missing rate case (**Table 3**), the results are quite similar to that in the low missing rate case. However, the 95% CPs are poor for the LOCF method in both moderate and large slope cases. In the high missing rate case (**Table 4**), the 95% CPs are very poor for the LOCF method in both moderate and large slope cases. In summary, the bias and RMSE will increase with the increase of the missing rate. However, the biases are small except for the LOCF method. For a fixed missing rate, the bias and RMSE will increase with the increase of the slope value.

Based on the six performance criteria for MCAR missing data, the LOCF is the poorest method. The CC method is the best method. However, the MI method had the same performance as the CC method. The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do NOT post your job titles, positions, academic degrees, zip codes, names of building/street/district/province/state, etc.). This template was designed for two affiliations.

Table 2. Performance of four methods for Case 1 and $\rho = 0.7$ (MCAR).

Slope (β_1)	Method	Bias (RMSE) of $\hat{\beta}_0$	Bias (RMSE) of $\hat{\beta}_1$	95% CP for $\hat{\beta}_0$	95% CP for $\hat{\beta}_1$
0.1	CC	**0.000 (**0.1251**)**	**0.000 (**0.0329**)**	89.5	92.9
	MS	**0.000** (0.1254**)**	**0.000 (0.0329)**	88.4	92.3
	LOCF	0.005 (**0.1237**)	−0.005 (**0.0322**)	**91.8**	**94.7**
	MI	**0.000** (0.1253)	**0.000** (0.0330)	90.1	93.2
2	CC	**0.000** (0.1248)	**0.000 (0.0316)**	89.0	**94.8**
	MS	**0.000** (0.1251)	**0.000 (0.0316)**	87.9	93.7
	LOCF	0.100 (0.1586)	−0.100 (0.1049)	**95.0**	40.5
	MI	**0.000 (0.1244)**	**0.000** (0.0317)	90.3	**95.2**
10	CC	**0.001** (0.1270)	**0.000** (0.0315)	89.1	**95.3**
	MS	**0.001** (0.1273)	**0.000 (0.0314)**	88.1	94.3
	LOCF	0.501 (0.5164)	−0.501 (0.5015)	**100.0**	0.0
	MI	**0.001 (0.1266)**	−0.001 (0.0315)	89.6	95.7

Table 3. Performance of four methods for Case 2 and $\rho = 0.7$ (MCAR).

Slope (β_1)	Method	Bias (RMSE) of $\hat{\beta}_0$	Bias (RMSE) of $\hat{\beta}_1$	95% CP for $\hat{\beta}_0$	95% CP for $\hat{\beta}_1$
0.1	CC	**0.001 (**0.1251**)**	**0.000** (0.0337)	88.4	93.3
	MS	**0.001** (0.1252)	**0.000** (0.0328)	84.9	90.9
	LOCF	0.018 (**0.1214**)	−0.013 (**0.0327**)	**94.6**	97.2
	MI	0.002 (0.1253)	−0.002 (0.0332)	90.8	**96.4**
2	CC	**−0.001** (0.1258)	**0.001** (0.0349)	88.0	92.2
	MS	**−0.001** (0.1257)	**0.001 (0.0340)**	85.5	89.6
	LOCF	0.351 (0.3708)	−0.251 (0.2525)	51.3	0.0
	MI	0.002 (**0.1254**)	−0.002 (0.0341)	**90.2**	**95.8**
10	CC	**0.000** (0.1287)	**0.000** (0.0348)	87.7	92.6
	MS	**0.000** (0.1286)	**0.000** (0.0338)	84.0	90.4
	LOCF	1.751 (1.7555)	−1.251 (1.2511)	0.1	0.0
	MI	0.003 (**0.1277**)	−0.002 (**0.0337**)	**90.0**	**95.0**

Table 4. Performance of four methods for Case 3 and $\rho = 0.7$ (MCAR).

Slope (β_1)	Method	Bias (RMSE) of $\hat{\beta}_0$	Bias (RMSE) of $\hat{\beta}_1$	95% CP for $\hat{\beta}_0$	95% CP for $\hat{\beta}_1$
0.1	CC	0.002 (0.1014)	**−0.001** (0.0336)	88.7	92.6
	MS	**0.001** (**0.0838**)	**−0.001** (**0.0253**)	81.6	86.5
	LOCF	0.038 (0.1217)	−0.027 (0.0367)	**94.9**	93.5
	MI	0.007 (0.1148)	−0.004 (0.0393)	91.9	**95.5**
2	CC	**0.000** (0.1016)	**0.000** (0.0337)	87.6	92.8
	MS	**0.000** (**0.0839**)	**0.000** (**0.0253**)	79.2	86.8
	LOCF	0.781 (0.2135)	−0.541 (0.0644)	0.1	0.0
	MI	0.006 (0.1149)	−0.003 (0.0394)	**91.5**	**96.0**
10	CC	**0.000** (0.1016)	**0.000** (0.0337)	89.0	93.5
	MS	**0.000** (**0.0839**)	**0.000** (**0.0253**)	81.8	87.0
	LOCF	3.901 (0.8862)	−2.700 (0.2672)	0.0	0.0
	MI	0.005 (0.1149)	−0.002 (0.0393)	**92.4**	**96.7**

4.2. Simulation Results for MAR Missing Data

Tables 5-7 show the simulation results for MAR (missing at random) data. The cases of low, moderate, and high missing rates are illustrated in **Table 5**, **Table 6**, and **Table 7**, respectively. In the low missing rate case (**Table 5**), the biases for the CC and MS methods are slightly larger for Intercept but smaller for those of Slope. However, the bias for the LOCF method is slightly smaller for Intercept but larger for that of Slope. Again, in the case that the slope is 10, the 95% CPs are very poor for the LOCF method (100% and 0% for Intercept and Slope). Also, in the case that the slope is 2, the 95% CP is 41.4% for the LOCF method. Based on the six performance criteria, the MI method is the best in this case. In the moderate missing rate case (**Table 6**), the results are similar for the performance of the LOCF method. Again, the MI method yield smaller bias, RMSE, and good 95% CP. In the high missing rate case (**Table 7**), the 95% CPs are very poor for the LOCF method in both moderate and large

Table 5. Performance of four methods for Case 1 and $\rho = 0.7$ (MAR).

Slope (β_1)	Method	Bias (RMSE) of $\hat{\beta}_0$	Bias (RMSE) of $\hat{\beta}_1$	95% CP for $\hat{\beta}_0$	95% CP for $\hat{\beta}_1$
0.1	CC	−0.029 (0.1253)	−0.001 (**0.0317**)	89.7	95.3
	MS	−0.029 (0.1255)	−0.001 (**0.0317**)	88.4	94.3
	LOCF	−0.006 (**0.1207**)	0.012 (0.0324)	**92.0**	**95.0**
	MI	**0.001 (**0.1222**)**	**−0.000 (**0.0319)	91.3	95.7
2	CC	−0.029 (0.1290)	−0.002 (**0.0328**)	88.7	93.9
	MS	−0.028 (0.1293)	**−0.001** (0.0330)	87.8	93.3
	LOCF	0.090 (0.1533)	−0.084 (0.0891)	**90.0**	41.4
	MI	**0.001 (0.1262)**	**−0.001 (**0.0329**)**	88.9	**94.5**
10	CC	−0.028 (0.1296)	−0.001 (**0.0326**)	89.0	93.8
	MS	−0.028 (0.1298)	−0.001 (**0.0326**)	88.0	93.2
	LOCF	0.490 (0.5054)	−0.483 (0.4842)	**100.0**	0.0
	MI	**0.002 (0.1267)**	**−0.000** (0.0327)	89.8	**94.2**

Table 6. Performance of four methods for Case 2 and $\rho = 0.7$ (MAR).

Slope (β_1)	Method	Bias (RMSE) of $\hat{\beta}_0$	Bias (RMSE) of $\hat{\beta}_1$	95% CP for $\hat{\beta}_0$	95% CP for $\hat{\beta}_1$
0.1	CC	0.032 (0.1247)	−0.050 (0.0605)	88.2	62.7
	MS	0.016 (**0.1216**)	−0.039 (0.0512)	86.6	68.3
	LOCF	−0.051 (0.1264)	0.043 (0.0518)	91.0	76.8
	MI	**0.001** (0.1219**)**	**−0.003 (0.0347)**	**91.4**	**96.0**
2	CC	0.033 (0.1276)	−0.050 (0.0599)	88.3	63.3
	MS	0.017 (**0.1246**)	−0.039 (0.0507)	85.7	68.3
	LOCF	0.281 (0.3054)	−0.195 (0.1966)	38.5	0.0
	MI	**0.001** (0.1267)	**−0.003 (0.0345)**	**90.2**	**94.5**
10	CC	0.032 (0.1235)	−0.050 (0.0602)	88.7	63.1
	MS	0.017 (**0.1204**)	−0.039 (0.0510)	86.8	68.5
	LOCF	1.682 (1.6854)	−1.195 (1.1949)	0.0	0.0
	MI	**−0.001 (**0.1213**)**	**−0.002 (0.0351)**	**91.5**	**94.5**

slope cases. In summary, the LOCF method yields larger biases, RMSEs, and poor 95% CPs in most cases. In contrast, the MI method performs much better than the other three methods under MAR mechanism.

5. Discussion and Conclusions

Although the simulation results suggested that the CC method was superior to the MS, LOCF, and MI methods under MCAR missing mechanism while MI method was superior to CC, MS, and MI methods under MAR, the performance of these methods actually depended on several factors especially the missing rate and time effect (that is, the size of the slope). However, there is no one single method that is the best under all situations.

Under the assumption of MCAR missing mechanism, when the missing rate increased from low to moderate (slope = 0.1 or 2), the values of estimated bias and RMSE for CC, MS, and MI methods were very close. Except

Table 7. Performance of four methods for Case 3 and $\rho = 0.7$ (MAR).

Slope (β_1)	Method	Bias (RMSE) of $\hat{\beta}_0$	Bias (RMSE) of $\hat{\beta}_1$	95% CP for $\hat{\beta}_0$	95% CP for $\hat{\beta}_1$
0.1	CC	0.095 (0.1575)	−0.104 (0.1097)	79.8	15.5
	MS	0.015 **(0.1256)**	−0.053 (0.0619)	82.6	48.4
	LOCF	−0.056 (0.1288)	0.052 (0.0581)	90.5	66.6
	MI	**0.008** (0.1348)	**−0.004 (0.0433)**	**92.5**	**96.5**
2	CC	0.093 (0.1572)	−0.103 (0.1094)	80.0	18.1
	MS	0.014 (**0.1269**)	−0.052 (0.0623)	81.7	48.9
	LOCF	0.684 (0.6934)	−0.461 (0.4619)	0.0	0.0
	MI	**0.006** (0.1392)	**−0.004 (0.0462)**	**91.4**	**95.2**
10	CC	0.096 (0.1590)	−0.104 (0.1101)	79.4	16.6
	MS	0.015 (**0.1262**)	−0.053 (0.0622)	84.0	47.2
	LOCF	3.805 (3.8064)	−2.621 (2.6212)	0.0	0.0
	MI	**0.009** (0.1367)	**−0.005 (0.0432)**	**93.0**	**96.9**

for high missing rate and large slope (that is, slope = 10), the values of bias and RMSE obtained by the MI method had large differences compared with those obtained by the CC and MS methods. This is not surprising at all because the CC method will yield unbiased estimated parameters under MCAR only with a small missing rate.

For the MAR missing data, the simulation results revealed that MI is the best method regardless of the missing rate and slope size based on bias, RMSE, and 95% CP. In fact, such a result is well documented in the literature [23] [31]. For low missing rate and small slope, the results did not differ significantly between the MI and LOCF methods. Such a result was also been discussed in the literature [39] [40].

In this paper, we consider a longitudinal study with five visiting time points and a total of 100 subjects. Three possible missing rates and three different slopes are used to mimic the real-world situations. In addition, two missing mechanisms are considered (that is, MCAR and MAR).Based on the simulation results, we have reached the following important conclusions: 1) CC method is the most appropriate method for handling MCAR missing data; 2) MI method is the most effective one in all simulated situations particularly under MAR setting because it yields smallest biases and has good 95% CP compared with the other methods; 3) the use of the LOCF method can potentially lead to imprecise parameter estimates hence can lead to invalid inferences.

In practice, inferior methods such as LOCF are still used for the longitudinal data analysis. The results via the simulation dataare indeed provide a good reference and rationale in choosing missing data handling method in order to obtain precise parameter estimates and valid inferences. Kenward and Molenberghs [41] did suggest that LOCF method should be avoided which is well supported by our simulation results.

Acknowledgements

I would like to thank the reviewer for his/her valuable comments and suggestions that make this paper much better in its contents.

References

[1] Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. John Wiley & Sons, New York.

[2] Collins, L.M., Schafer, J.L. and Kam, C.M. (2001) A Comparison of Inclusive and Restrictive Missing-Data Strategies in Modern Missing-Data Procedures. *Psychological Methods*, **6**, 330-351. http://dx.doi.org/10.1037/1082-989X.6.4.330

[3] Little, R.J.A. (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, **83**, 1198-1202. http://dx.doi.org/10.1080/01621459.1988.10478722

[4] Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002) Analysis of Longitudinal Data. 2nd Edition, Clarendon Press, Clarendon.

[5] Carpenter, J.R., Kenward, M.G. and Vansteelandt, S. (2006) A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data. *Journal of the Royal Statistical Society, Series A* (*Statistics in Society*), **169**, 571-584. http://dx.doi.org/10.1111/j.1467-985X.2006.00407.x

[6] Musil, C.M., Warner, C.B., Yobas, P.K. and Jones, S.L. (2002) A Comparison of Imputation Techniques for Handling Missing Data. *Western Journal of Nursing Research*, **24**, 815-829. http://dx.doi.org/10.1177/019394502762477004

[7] Sprint, A. and Dupin-Sprint, T. (1993) Imperfect Data Analysis. *Drug Information Journal*, **27**, 995-994.

[8] Myers, W.R. (2000) Handling Missing Data in Clinical Trials: An Overview. *Drug Information Journal*, **34**, 525-533.

[9] Hening, D. and Koonce, D.A. (2014) Missing Data Imputation Method Comparison in Ohio University Student Retention Database. *Proceeding of the* 2014 *International Conference on Industrial Engineering and Operations Management*, Bali, Indonesia.

[10] Ali, A.M.G., Dawson, S.J., Blows, F.M., Provenzano, E., Ellis, I.O., Baglietto, L., Huntsman, D., Caldas, C. and Pharoah, P.D. (2011) Comparison of Methods for Handling Missing Data on Immunohistochemical Markers in Survival Analysis of Breast Cancer. *British Journal of Cancer*, **104**, 693-699.

[11] Patrician, P.A. (2002) Focus on Research Methods Multiple Imputation for Missing Data. *Research in Nursing & Health*, **25**, 76-84. http://dx.doi.org/10.1002/nur.10015

[12] Nakai, M., Chen, D.G., Nishimura, K. and Miyamoto, Y. (2014) Comparative Study of Four Methods in Missing Value Imputations under Missing Completely at Random Mechanism. *Open Journal of Statistics*, **4**, 27-37.

[13] Lavori, P.W., Dawson, R. and Shera, D. (1995) A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data. *Statistics in Medicine*, **14**, 1913-1925. http://dx.doi.org/10.1002/sim.4780141707

[14] Allison, P.D. (2001) Missing Data. Sage Publications, Thousand Oaks.

[15] Kim, J.O. and Curry, J. (1977) The Treatment of Missing Data in Multivariate Analysis. *Sociological Methods Research*, **6**, 215-240. http://dx.doi.org/10.1177/004912417700600206

[16] Allison, P.D. (1998) Multiple Regression: A Primer. Pine Forge Press, Thousand Oaks.

[17] Little, R.J.A. (1992) Regression with Missing X's: A Review. *Journal of the American Statistical Association*, **87**, 1227-1237.

[18] Greenland, S. and Finkle, W.D. (1995) A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology*, **142**, 1255-1264.

[19] Schafer, J.L. and Graham, J.W. (2002) Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**, 147-177. http://dx.doi.org/10.1037/1082-989X.7.2.147

[20] Carpenter, J., Kenward, M.G., Evans, S. and White, I. (2004) Last Observation Carry-Forward and Last Observation Analysis. *Statistics in Medicine*, **23**, 3241-3242. http://dx.doi.org/10.1002/sim.1891

[21] Cook, R.J., Zeng, L.L. and Yi, G.Y. (2004) Marginal Analysis of Incomplete Longitudinal Binary Data: A Cautionary Note on LOCF Imputation. *Biometrics*, **60**, 820-828. http://dx.doi.org/10.1111/j.0006-341X.2004.00234.x

[22] Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2006) Analyzing Incomplete Discrete Longitudinal Clinical Trial Data. *Statistical Science*, **21**, 52-69. http://dx.doi.org/10.1214/088342305000000322

[23] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York. http://dx.doi.org/10.1002/9780470316696

[24] Tabachnick, B.G. and Fidell, L.S. (2000) Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC, Boca Raton.

[25] Molenberghs, G., Thijs, H., Jansen, I., *et al*. (2004) Analyzing Incomplete Longitudinal Clinical Trial Data. *Biostatistics*, **5**, 445-464. http://dx.doi.org/10.1093/biostatistics/kxh001

[26] Shao, J. and Zhong, B. (2003) Last Observation Carry-Forward and Last Observation Analysis. *Statistics in Medicine*, **22**, 2429-2441. http://dx.doi.org/10.1002/sim.1519

[27] Mallinckrodt, C.H., Clark, W.S. and David, S.R. (2001) Accounting for Dropout Bias Using Mixed-Effects Models. *Journal of Biopharmaceutical Statistics*, **11**, 9-21. http://dx.doi.org/10.1081/BIP-100104194

[28] Mallinckrodt, C.H., Kaiser, C.J., Watkin, J.G., Detke, M.J., Molenberghs, G. and Carroll, R.J. (2004) Type I Error Rates from Likelihood-Based Repeated Measures Analyses of Incomplete Longitudinal Data. *Pharmaceutical Statistics*, **3**, 171-186. http://dx.doi.org/10.1002/pst.131

[29] Gadbury, G.L., Coffey, C.S. and Allison, D.B. (2003) Modern Statistical Methods for Handling Missing Repeated Measurements in Obesity Trials: Beyond LOCF. *Obesity Reviews*, **4**, 175-184. http://dx.doi.org/10.1046/j.1467-789X.2003.00109.x

[30] Rubin, D.B. (1977) Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, **72**, 538-543. http://dx.doi.org/10.1080/01621459.1977.10480610

[31] Schafer, J.L. (1997) The Analysis of Incomplete Multivariate Data. Chapman & Hall, London. http://dx.doi.org/10.1201/9781439821862

[32] Schafer, J.L. (2000) Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC, Boca Raton.

[33] Rubin, D.B. (2004) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York.

[34] Bodner, T.E. (2008) What Improves with Increased Missing Data Imputations? *Structural Equation Modeling*: *A Multidisciplinary Journal*, **15**, 651-675. http://dx.doi.org/10.1080/10705510802339072

[35] Dmitrienko, A., Molenberghs, G., Chuang-Stein, C. and Offen, W. (2005) Analysis of Clinical Trials Using SAS: A Practical Guide. SAS Institute Inc., Cary.

[36] Yuan, Y.C. (2000) Multiple Imputation for Missing Data: Concepts and New Development. SAS Institute Inc., Rockville.

[37] Allison, P.D. (2000) Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods and Research*, **28**, 301-309. http://dx.doi.org/10.1177/0049124100028003003

[38] Huang, R. and Carriere, K.C. (2006) Comparison of Methods for Incomplete Repeated Measures Data Analysis in Small Samples. *Journal of Statistical Planning and Inference*, **136**, 235-247. http://dx.doi.org/10.1016/j.jspi.2004.06.005

[39] Unnebrink, K. and Windeler, J. (2001) Intention-to-Treat: Methods for Dealing with Missing Values in Clinical Trials of Progressively Deteriorating Diseases. *Statistics in Medicine*, **20**, 3931-3946. http://dx.doi.org/10.1002/sim.1149

[40] Halabi, S., Wun, C.C. and Davis, B.R. (2003) Analysis of Survival Data with Missing Measurements of a Time-Dependent Binary Covariate. *Journal of Biopharmaceutical Statistics*, **13**, 253-270. http://dx.doi.org/10.1081/BIP-120019270

[41] Kenward, M.G. and Molenberghs, G. (2009) Last Observation Carried Forward: A Crystal Ball? *Journal of Biopharmaceutical Statistics*, **19**, 872-888. http://dx.doi.org/10.1080/10543400903105406

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.