Research Article
Open Access
CrossMark

Missing data analysis for binary multivariate longitudinal data through a simulation study

Abstract

Longitudinal data play an important role in many biomedical and social problems. Multivariate longitudinal data is a generalization of univariate longitudinal data where there are many dependent outcomes obtained at many occasions. Missing data frequently occur in longitudinal studies for many reasons such as subjects' moving or medical issues. Missingness could affect the bias and precision of parameter estimations. In this paper, we fill this gap with a dedicated simulation to study the missingness effects on the parameter estimations for multivariate longitudinal data. The simulation study is conducted to evaluate the effects of the correlation in the parameter estimations for missing data mechanisms, that are MCAR, MAR and MNAR. Also, several analysis methods for handling missing data are used to reduce the bias of the parameter estimations when there is missingness in the responses, namely Completed cases (CC), Mean substitution (MS), Last Observation Carried Forward (LOCF) and Regression Imputation (RI). The inverse propensity weighed (IPW) GEE is used for the MAR mechanisms to investigate its performance with dropout missingness. Based on evaluation measurements such as root mean square error (RMSE) and coverage probability (CP), we found our results in agreement with Little[1] who indicated MCAR is the appropriate mechanism for the longitudinal data. We found that severe correlation over the occasions may affect the parameter estimations for both complete and incomplete data. Also, we found the MS handling method, after extension to accommodate the multivariate structure, largely has good results even when one outcome is MAR and the second is MCAR. The inverse propensity weighed GEE shows some good results to treat the dropout in MAR mechanism especially when the correlation is induced over the outcomes.

Volume 7 Issue 2 - 2018

Hissah Alzahrani
Umm Al-Qura University, Saudi arabia

Correspondence: Hissah Alzahrani, Mathematical Science Department, College of applied sciences, Umm Al-Qura University, Mekkah, 24382, Saudi Arabia,
Email hahzahrani@uqu.edu.com.sa

Received: February 02, 2018 | **Published:** March 12, 2018

Introduction

The nature of longitudinal data is repeated measurements over many occasions. Missingness frequently occurs during a longitudinal study because of circumstances such as a subject moving, medical illness or administrative reasons. In this paper, our focus is multivariate longitudinal data, for which there are more than one outcome measured at many occasions. The missingness in multivariate longitudinal data can happen such that all outcomes are missing at a particular time or partially such that only a subset of outcomes is not observed at that time. Missingness is considered a big issue in statistical analysis since its effects are to reduce the statistical power and increase bias. The statistical analysis should pay attention to the problem of missing data and use some statistical missing data handling methods to reduce the bias and reach better estimates. Missing data mechanisms were classified by Little[1] into three types:

a. Missing completely at random (MCAR), when the missingness is independent of the observed and unobserved responses;

b. Missing at random (MAR), when the missingness may depend on observed responses, but is independent of the unobserved responses;

c. Missing not at random (MNAR), when missingness depends on both observed and unobserved responses. MNAR is often called "informative missingness." In real life, the missingness mechanism is unknown and it is not a trivial issue to assume that missingness is MCAR or MAR. It is important to assume the appropriate mechanism based on the nature of the study to avoid biased estimates. In this paper, we generated artificial data through a simulation study to perform statistical analysis of incomplete binary multivariate longitudinal data in order to answer many questions. First, are the estimated correlation parameters different for the completed data and incomplete data, controlling the correlation over the outcomes and occasions? Second, are the estimated regression parameters valid for the three cases of missingness in all outcomes, in just partial outcomes, or mixed missing mechanisms for the outcomes. The third question is about the best method for handling missing data among CC, MS, LOCF and RI to treat the incomplete binary multivariate longitudinal data.

Many researchers have conducted simulation studies to investigate the effects of missing data handling methods for univariate longitudinal data. For example Myers[2] conducted a simulation for longitudinal data to compare the complete case method and multiple imputation methods in clinical trials and found some limitations of using multiple imputation. Also, Touloumi et al.[3] designed a simulation study to investigate the impact of missing data due to drop out in longitudinal studies for six methods such as GEE, weighted and unweighted ordinary least square. They found the MCAR assumption is doing well for all their methods while MAR and MNAR generate biased estimates. Newman[4] compared six missing data techniques based on simulation study. The comparison was for the three mechanisms of the missingness (MCAR, MAR, and MNAR) and for three levels of missingness 25%, 50% and 75%. Their results support a multiple imputation approach. Also, Hening[5] performed a comparison for five imputation methods for the missing data (mean substitution, median substitution, zero values, hot deck and MI) with 20% missing rates for the first year students retention data in Ohio university. His

dissertation indicates that mean imputation and median imputation yield good performance in precision. Ali et al.[6] investigated four imputation methods (complete case analysis, mean substitution, and multiple imputation with and without inclusion of the outcome in the imputation model) under MCAR, MAR and MNAR based on simulation study to figure out the best approach. They found, based on their model, that the estimates for multiple imputation were least biased and most accurate.

The purpose of this paper is an investigation of the effect of missing data and the performance of missing data handling methods, especially for binary multivariate longitudinal data through a simulation study. Multivariate binary longitudinal data are generated for specified correlation structures and for different missing data mechanisms. The organization of this paper is as follows: section 3 contains the simulation design. It contains the model details, correlation scenarios, missing data mechanisms, handling missing data methods and simulation evaluation measurements. In section 4, the simulation results are produced. Finally, discussion and concluding remarks are in section 5.

Simulation design

The multivariate longitudinal data structure is an extension of the univariate longitudinal structure to more than one outcome. Each individual i has a vector of responses for different outcomes, $j=1,2\ldots,J_i$. Also, each individual is measured at different occasions, $j=1,2\ldots,J_i$, and has cluster size $n_i=J_iK$. To simplify these notations, we will refer to J as J, which is the number of occasions for all individuals. The vector of completed responses for subject i is:

$$Y_i=[Y_{i11},Y_{i21},\ldots,Y_{iJ1},Y_{i12},Y_{i22},Y_{i32},\ldots,Y_{iJ2},\ldots\ldots\ldots\ldots,Y_{i1K},Y_{i2K},\ldots,Y_{iJK}]^T.$$

We conducted a simulation study in order to explore the changes of the parameter estimations for different missing data mechanisms when the data structure is multivariate longitudinal binary data. One of the goals to design the simulation study was to control the missing data pattern and the correlation for multivariate structure through the regression model.

Simulation model

We will use generalized estimating equations (GEE), an extension of generalized linear modeling to longitudinal data, to specify the simulation model and for the analysis. Thus, we specify a marginal model for the correlated responses. Estimation via GEE yields consistent regression parameter estimations despite the lack of full likelihood specification when the data are complete. This consistency also holds when data are MCAR, as shown in Little.[1] We fit the GEE model of Shelton et al.[7] to estimate the effects for each outcomes separately using a Kronecker product approach to account for the correlation structure. Let $X_i=0,1$ be the treatment assignment covariate. The time covariate is $t_j=1,2,3$ for three occasions. Let Y_{ijk} be the binary response that is measured at time $j, j=1,2,3$ for observation $i, i=1,2,3,\ldots,N$ and for outcome $k=1,2$. Then, we assume the logistic model

$$logit(E(Y_{ijk}|X_i))=\beta_{0k}+\beta_{1k}X_i+\beta_{2k}t_j$$

with the true parameter values $\beta_{01}=0.2$, $\beta_{02}=0.1$, $\beta_{11}=0.05$, $\beta_{12}=-0.15$, $\beta_{21}=0.05$, and $\beta_{22}=-0.25$. Given the vector of treatment covariate X_i, the i for subject i is assumed to follow the Bernoulli distribution, $Y_{ijk}|X_i\sim Ber(E(Y_{ijk}|X_i))$. We did the simulation for the marginal regression model based on specified correlation structures and for different missing data mechanisms. We applied the method of Alzahrani[8] that used the bridge distribution for the random effect in the mixed model to generate the correlated binary data. We generated $N=250$ samples of correlated binary data, conducted the simulation study for eight missing data mechanisms and for five scenarios (correlation structures) to explore the properties of the model when the correlation is induced over the outcomes and the occasions. In the following subsections, there are more explanations about correlation scenarios, the missing data mechanisms, handling missing data methods and evaluation measurements.

Correlation scenarios

Multivariate longitudinal data has a complicated correlation structure in which the correlation has two factors, over the occasions and over the outcomes. This simulation study explores the effects of increasing correlation over each of these two factors. We start by explaining the nature of the correlation in multivariate longitudinal data. The correlation matrix $R(\gamma)$ is a function of $\gamma_{jk,j'k'}$, where $\{\gamma_{jk,j'k'}\}$ represents the collection of within subject correlation parameters of size $\binom{JK}{2}$ of all non-redundant pairwise correlation parameters. For our setup with $J=3$ and $K=2$, there are 15 correlation parameters given by:

$$\gamma_{jk,j'k'}=\frac{P(Y_{jk}=1,Y_{j'k'}=1)-P(Y_{jk}=1)P(Y_{j'k'}=1)}{\sqrt{P(Y_{jk}=1)P(Y_{j'k'}=1)(1-P(Y_{jk}=1))(1-P(Y_{j'k'}=1))}}$$

$$R=Y_{11}Y_{21}Y_{31}Y_{12}Y_{22}Y_{32}Y_{11}1\gamma_{11,21}\gamma_{11,31}\gamma_{11,12}\gamma_{11,22}\gamma_{11,32}Y_{21}-1$$

$$\gamma_{21,31}\gamma_{21,12}\gamma_{21,22}\gamma_{21,32}Y_{31}-1\gamma_{31,12}\gamma_{31,22}\gamma_{31,32}Y_{12}---1\gamma_{12,22}\gamma_{12,32}Y_{22}----1\gamma_{22,32}Y_{32}-----1$$

To reduce size of the unknown correlation parameter vector γ, we assume there are three components that build up the correlation structure R in the multivariate longitudinal data. The inter-outcome (υ), the intra-outcome (α); and the cross association (τ). The αs represent the correlation between the outcomes in the same occasion. The υs represent the correlation within outcomes at different occasions, and τs represent the correlation between different outcomes measured at different occasions. Thus, R can be written as follows:

$$R=Y_{11}Y_{12}Y_{13}Y_{21}Y_{22}Y_{23}Y_{11}1\upsilon_1\upsilon_2\alpha_1\tau_1\tau_2Y_{12}-1\upsilon_3\tau_3\alpha_2\tau_4Y_{13}--1\tau_5\tau_6\alpha_3Y_{21}---1\upsilon_4\upsilon_5Y_{22}----1\upsilon_6Y_{23}-----1$$

We conducted five scenarios under the assumption of exchangeability for each correlation type υ,α and $_{\tau=0}$ and $\tau=0$ for all scenarios. We faced a difficulty to do the rest of the scenarios to the restrictions on generating correlated binary outcomes from the regression model for specified correlation strictures. The following table shows the values for each correlation parameter in each scenario for the correlation matrix R (Table 1).

Missing data mechanisms

It is important to investigate the nature of the missing data to get

Citation: Alzahrani H. Missing data analysis for binary multivariate longitudinal data through a simulation study. *Biom Biostat Int J.* 2018;7(2):103–113. DOI: 10.15406/bbij.2018.07.00196

valid inference. The missing data mechanism can be defined as the probability distribution of the missing indicator variable R_i=(0,1) that takes value 1 when the response is missing, and 0 if not. The vector of complete responses for subject i is:

$$Y_i = [Y_{i11}, Y_{i21}, \ldots, Y_{iJ1}, Y_{i12}, Y_{i22}, Y_{i32}, \ldots, Y_{iJ2}, \ldots\ldots\ldots\ldots, Y_{i1K}, Y_{i2K}, \ldots, Y_{iJK}]^T$$

Let R_i be the vector of response missingness indicators,

$$R_i = [R_{i11}, R_{i21}, \ldots, R_{iJ1}, R_{i12}, R_{i22}, R_{i32}, \ldots, R_{iJ2}, \ldots\ldots\ldots\ldots, R_{i1K}, R_{i2K}, \ldots, R_{iJK}]^T,$$

with R_{ijk}=1 when Y_{ijk} is not observed, and R_{ijk}=0 when R_i is observed. In this paper, we do not consider missingness in the covariates. Given R_i, the complete set of responses Y_i can be divided into two groups: Y_i^M and Y_i^M. Y_i^O denotes the vector of observed responses and Y_i^M the vector of missing responses. Setting up the three mechanisms is done by the approaches in Gibbons.[9] Assuming the responses at the first time are fully observed, then the three missing mechanisms are as follows.

Table 1 The scenarios of correlation design

	$\alpha = 0.00$	$\alpha = 0.60$	$\alpha = 0.90$
$\upsilon = 0.00$	Scenario 1	Scenario 2	Scenario 3
$\upsilon = 0.90$	Scenario 4	-	-
$\upsilon = 0.90$	Scenario 5	-	-

Missing completely at random MCAR:

The missing data pattern is considered to be MCAR if the probability that the responses are missing is independent of both Y_i^O and Y_i^M, $P(R_i|Y_i)=P(R_i)$. The missing data can be missing arbitrary or non-arbitrary. Here for MCAR mechanism, we will set up arbitrary missingness. In the simulation design, we assume the missingness is 25% at time 2 and 25% at time 3. Because we generate drop out missingness for the MAR and MNAR mechanisms, we checked that the drop out patterns in our MCAR data satisfied MCAR. Let D_i=0 for subject i who has data at all time points t=1,2,3 (no drop out), D_i=2 for the dropout case at time 1 and D_i=2 when the dropout occurs at time 2. Also, let $last$ denote the last observed value of the response for subject i. Then, a logistic regression model for each outcome is done separately as follows:

$$log[\frac{P(D_i=j|D_i\geq j)}{1-P(D_i=j|D_i\geq j)}] = \beta_{0j} + \beta_1 last_i + \beta_2 X_i + \beta_3 last_i t_j$$

The MCAR mechanism implies that β_3 and β_3 are zero. Hence, a test for MCAR is as test of the null hypothesis H_0:β_1=β_3=0, which is not rejected when $\hat{\beta}_1$ and $\hat{\beta}_3$ are both not statistically different from zero. If the simulated dataset led to rejection of H_0, we discarded those data and generated a new data set until H_0 was not rejected.

Missing at random MAR

The missing at random mechanism implies the missingness probability is independent of Y_i^M, i.e., $P(R_i|Y_i)=P(R|Y_i^O)$. Here the design is much easier than for MCAR. We just design the missingness to be related the observed data and independent of unobserved data. In this simulation design, we set up the drop out after the first time point. If the logit of the response is lower than a specified cutpoint, then the subject drops out at the next time point for all the subsequent times.

Missing not at random MNAR

The missing not at random mechanism holds when the probability the responses are missing is not independent of either Y_i^M or Y_i^O. Then we set the missingness related to the observed and unobserved data. After the first time point, if the logit of the response is lower than a specified cutpoint, then the subject drops out at that time point for all the subsequent times. Here the cause of the missingness comes from observed and unobserved data, satisfying MNAR. To address the multivariate structure, we further classified mechanisms as "both missingness" when missingness occurs in both outcomes, and "partial missingness" when only one outcome is incomplete. The term "mixed missingness" when the two outcomes have different missing mechanisms. Defining different missing patterns is in order to study the changes in the regression effects estimations and the correlation parameters. In the following table, there are eight missing data mechanisms (Figure 1) (Table 2).

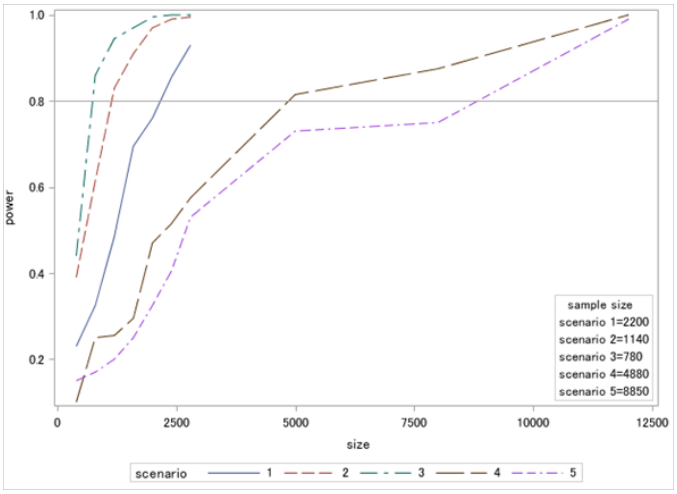

Figure 1 Test power versus sample size.

Table 2 The mechanisms of incomplete-dataset

Type	Outcome1	Outcome2	Code
Both	MCAR	MCAR	MCAR_MCAR
Both	MCAR	MCAR	MCAR_MCAR
Both	MNAR	MNAR	MNAR_MNAR
Partial	Completed	MCAR	COMP_MCAR
Partial	Completed	MAR	COMP_MAR
Partial	Completed	MNAR	COMP_MNAR
Mixed	MCAR	MAR	MCAR_MAR
Mixed	MCAR	MNAR	MCAR_MNAR

Handling missing data

There are many traditional methods of handling missing data in longitudinal studies. We conducted four methods as follows.

Citation: Alzahrani H. Missing data analysis for binary multivariate longitudinal data through a simulation study. *Biom Biostat Int J.* 2018;7(2):103–113. DOI: 10.15406/bbij.2018.07.00196

Completed case analysis

This method is defined by deleting all subjects who have missing data. It includes only the completed cases. The primary condition to use this method for valid inference is the MCAR assumption. Based on Little,[1] It will yield unbiased estimates of the mean responses when the assumption of the missing data is MCAR. However, the completed case analysis is not appealing when the size of the completed subjects is small relative to the whole number of subjects because it will reduce the statistical power and increase the standard error of the estimates Allison.[8] Also, if the data is not MCAR, then the parameter estimates could be biased.

Imputation

The imputation approach is widely used in practice. The basic idea is filling the missing values with imputed values. The imputed values are chosen by many methods. We will discuss three methods: Mean Substitution (MS), Last Observation Carried Forward (LOCF) and Regression Imputation (RI). The advantage of using imputation methods is the possibility to use the standard statistical analysis methods for the completed data after the imputation.

Mean substitution (MS)

In this method, we easily use single imputation for each missing data by the mean of available observed values. This method may causes bias in the parameter estimations and under estimate the variability Cook et al.[10] Thus, it may artificially reduce the variability of the parameter estimations. In this paper, we accommodate the multivariate structure by imputing the mean of each outcome, for each occasion, and for the same treatment. That means when we have two outcomes, three occasions and treatment binary variable, we will impute the missing values by 12 means instead of single mean. Each missing value is imputed by the mean of its outcome, its occasion, and treatment.

Last observation carried forward (LOCF)

The LOCF is frequently used in clinical trials, it is a single imputation method that is widely used for longitudinal data, despite that it leads to bias expect for MCAR mechanism. In this method, each missing value is replaced by the last observed value for that subject. It is easy to impute LOCF method, but it is unrealistic to assume the observations following the dropouts remain unchanged unless the dropout is due to cure. Shao & Zhong[11] conducted studies and proved that LOCF method could have biased results and reduce the precision of the parameter estimations.

Regression imputation (RI)

The idea behind the regression imputation is simple. It is based on assuming the distribution of the missing responses is close to the distribution of the observed responses. Here, we impute the missing responses using the prediction from the regression equation derived from the observed responses. Then, we remodel the completed responses after the imputation, using the prediction method to impute the missing responses seems not appealing when the cause of the drop out is related to the efficiency of the treatment. For example, when the sick subjects tends to drop out more than the healthy subjects, this leads to different estimates of the observed and missing responses. In MAR and MNAR assumptions, the there is a statistical difference between responses means of the observed and missing responses.

Inverse probability weight method (IPW)

The GEE method can be adapted to treat the dropout missingness for MAR mechanism. One applicable method is the inverse probability weighted (IPW) GEE approach of Preisser et al.[12] and Robins & Rotnitzky.[13] It provides a method to handle the dropout missingness in the lack of good reasons to assume the MCAR mechanism. The idea is to weight the observed data to account for the probability of unobserved data. The weights are obtained by estimating the probability of dropout as a function of observed responses prior to dropouts and covariates. The probability of not dropping out is called the propensity score. Here in the multivariate longitudinal data, the appropriate way is to estimate the propensity score weights for each outcome k. Let π_{ijk} denote the conditional probability of subject i being observed at time j for outcome k given the readings history of that subject at prior times.

$$\pi_{ijk}=P(R_{ijk}=0\,|R_{i1k}=R_{i2k}=\ldots=R_{i,j-1,k},X_i,Y_{i1k},Y_{i2k},\ldots,Y_{i,j-1,k})$$

The appropriate weight for Y_{ijk} is the inverse of the unconditional probability of Y_{ijk} being observed, this probability being the cumulative product of π_{ijk} for : j=1,2,..,J

$$w_{ijk}=\frac{1}{\pi_{i1k}\pi_{i2k}\ldots\pi_{iJk}}$$

In our simulation context, the model for missingness estimates the probability of not dropping out at given time point, given the previous response, treatment values and their interaction. The logistic regression model for each outcome k is:

$$logit(\pi_{ijk})=\beta_{0k}+\beta_{1k}t_{ij}+\beta_{3k}X_i+\beta_{4k}Y_{i,j-1,k}+\beta_{5k}X_iY_{i,j-1,k}.$$

Thus score propensity is estimated for each subject at each time. Then, we analyze the data using weighted GEE model of Shelton[7] using the "working independence" correlation matrix.

Simulation evaluation

We used five criteria to measure the performance of each scenario and mechanism for the correlated binary data. These are the average, standard deviation (std), the bias to express the difference between the true and estimated parameters, the root of mean square error (RMSE), and the coverage probability (CP) associated with the usual 95% confidence interval. Using the RMSE is good because it is the square root of MSE that measures the precision of the estimates. The coverage probability is the proportion of the nominal 95% confidence intervals from simulated datasets that contain the true parameter values.

The results

The general strategy of this study is generating the artificial multivariate longitudinal data for two outcomes and three occasions, and five correlation scenarios based on the model in 2.1. Then, we set up the eight missing mechanisms described in 2.3. First, The incomplete datasets are analyzed, and the parameters are estimated. The correlation parameters are estimated using sample correlation values for each mechanism, and for each correlation scenario. In scenario 1, where all correlation parameters are assumed to be zero, the estimated parameters are close to zero.

In scenarios 2 and 3, we induced the correlation between the outcomes in times 1, 2 and 3 respectively, α_1, α_2 and α_3. Table 3 reports the estimated means of 224 samples of the outcomes parameters. In scenario 2 and scenario 3, the true correlation parameters are 0.6 and 0.9, respectively. The estimated correlation

Citation: Alzahrani H. Missing data analysis for binary multivariate longitudinal data through a simulation study. *Biom Biostat Int J.* 2018;7(2):103–113. DOI: 10.15406/bbij.2018.07.00196

parameters in the completed data exhibit increasing bias over time, and this may be due to existence of the time covariate. Comparing the estimated correlation parameters for the complete and incomplete data for different mechanisms, we notice many points. We found α_1 did not change for all the missing mechanisms because the baseline is full without missingness. The MNAR assumption has a clear effect on α_2 and α_3 over the both, partial and mixed mechanisms. In Table 4, the estimated correlation parameters for time factor υ_1, υ_3 and υ_3 represent the correlation between the responses at (time1, time2), (time1, time3) and (time2, time3) within the first outcome and υ_4, υ_6 and υ_6 within the second outcome. In scenarios 4 and 5, we induced the correlation over the occasions factor. We found the MCAR_MCAR and COMP_MCAR yield results very close to parameters in the completed data but the MCAR_MAR has been affected in scenario 5, the scenario that has strongest correlation over the occasions.

Table 3 The within outcomes correlation parameter estimations in scenario 2 and 3

	Scenario 2			**Senario 3**		
	α_1	α_2	α_3	α_1	α_2	α_3
Complete	0.572	0.538	0.496	0.772	0.668	0.573
MCAR_MCAR	0.572	0.539	0.495	0.772	0.668	0.572
MAR_MAR	0.572	0.539	0.496	0.772	0.668	0.576
MNAR_MNAR	0.572	0.372	0.197	0.772	0.512	0.124
COMP_MCAR	0.572	0.539	0.495	0.772	0.668	0.572
COMP_MAR	0.572	0.538	0.495	0.772	0.668	0.576
COMP_MNAR	0.572	0.404	0.254	0.772	0.512	0.124
MCAR_MAR	0.572	0.539	0.495	0.772	0.668	0.576
MCAR_MNAR	0.572	0.405	0.253	0.772	0.512	0.142

Additionally, the results of the remaining MAR and MNAR for both, partial and mixed missingness are biased and have been affected by the correlation apart from the baseline parameters α_1 and α_4. We found the MCAR assumption for both outcomes is the appropriate assumption for the GEE model. Now we can figure out its advantage to keep the correlation parameter close to the correlation parameters in the complete data. Also, we may find the missingness pattern of MAR and MNAR mechanisms could affect the parameter estimations due to the bias in the correlation parameters before we fit the model.

In model 1, there are three regression coefficients associated with the intercept, X and time covariates for each outcome after we fit the GEE model assuming the unstructured within subject correlation. The parameter estimations of the X covariate are β_{11},β_{12} for scenarios 1, 2 and 4 in Tables 5–7, respectively. Generally from all the scenarios, we found the results of MCAR_MCAR and COMP_MCAR mechanisms are close to the regression coefficients of the completed data without missngness. The MAR_MAR and MNAR_MNAR have shown some bias results. In scenarios 2 and 4, the bias clearly affects the MAR more than MNAR in the parameters of the complete outcome. For the partial mechanisms, we found in scenario 1 that the estimated effects appear unbiased for the for the first outcome, which is complete. We conclude that assuming MCAR, which is implicit for GEE models, is affected by MAR and MNAR when they are mixed, especially with the strong correlation over the occasions. Now, we present the analysis after handling the missing data mechanisms for the incomplete data. We handled the missing data using four different methods: CC, MS, LOCF and RI as they are described in section 2.4. In Table 8 the results of the X covariate parameters estimations after handling the missingness just for the mechanism MCAR_MCAR while the rest of the intercepts and time covariate are in appendix B. Since the appropriate mechanism for the GEE models is the assumption MCAR, we will analyze the handling methods for the mechanisms COMP_MCAR, MCAR_MCAR, COMP_MAR and MCAR_MAR. In Table 8 the results of the estimated X effects on the log odds of response $P(Y_{ijk}=1)$ for the MCAR_MCAR mechanism. We present four evaluation measurements. We found using the MS and LOCF methods have good means, CP and also less bias in scenario 1 for the estimated parameters of the treatment effects. In scenario 2 where the correlation is induced between the outcomes, we found the methods LOCF and MS have better means close to the true values, good CP and less bias. For the results of scenario 4 where the correlation is induced over the occasions, we found the parameter estimation of the completed data are already biased and we don't find good method to reduce it over the four methods.

Table 4 The within occasions correlation parameter estimations in scenario 4 and 5

Senario 4							Senario 5					
Within outcome 1			**Within outcome 2**				**Within outcome 1**			**Within outcome 2**		
Complete	0.592	0.589	0.587	0.584	0.567	0.581	0.883	0.876	0.878	0.852	0.773	0.849
MCAR_MCAR	0.59	0.589	0.589	0.583	0.568	0.583	0.883	0.877	0.881	0.852	0.772	0.849
MAR_MAR	0.497	0.371	0.263	0.496	0.371	0.275	0.844	0.642	0.621	0.813	0.442	0.626
MNAR_MNAR	0.445	0.324	0.303	0.423	0.274	0.291	0.811	0.746	0.746	0.743	0.447	0.645
COMP_MCAR	0.59	0.589	0.589	0.584	0.567	0.581	0.883	0.877	0.881	0.852	0.773	0.849
COMP_MAR	0.59	0.589	0.589	0.496	0.371	0.275	0.883	0.877	0.881	0.813	0.442	0.626
COMP_MNAR	0.59	0.589	0.589	0.423	0.274	0.291	0.883	0.877	0.881	0.743	0.447	0.645
MCAR_MAR	0.592	0.589	0.587	0.496	0.371	0.275	0.883	0.876	0.878	0.813	0.442	0.626
MCAR_MNAR	0.592	0.589	0.587	0.423	0.274	0.291	0.883	0.876	0.878	0.743	0.447	0.645

Citation: Alzahrani H. Missing data analysis for binary multivariate longitudinal data through a simulation study. *Biom Biostat Int J.* 2018;7(2):103–113. DOI: 10.15406/bbij.2018.07.00196

Table 5 The estimates of X covariate in scenario 1 for the two outcomes

	True $\beta_{11} = 0.05$					True $\beta_{12} = -0.15$				
Type	Mean	Std	Bias	RMSE	CP	Mean	Std	Bias	RMSE	CP
Complete	0.055	0.112	0.005	0.112	94.2	-0.148	0.103	0.002	0.103	95.98
MCAR_MCAR	0.059	0.115	0.009	0.116	96.88	-0.145	0.115	0.005	0.114	95.54
MAR_MAR	0.055	0.121	0.005	0.121	92.92	-0.153	0.121	-0.003	0.121	94.81
MNAR_MNAR	0.049	0.13	-0.001	0.13	94.2	-0.13	0.125	0.02	0.126	94.2
COMP_MCAR	0.056	0.112	0.006	0.112	94.2	-0.145	0.115	0.005	0.114	95.98
COMP_MAR	0.055	0.111	0.005	0.111	93.84	-0.15	0.121	0	0.121	94.79
COMP_MNAR	0.055	0.112	0.005	0.112	93.75	-0.13	0.125	α_2 0.02	0.126	94.2
MCAR_MAR	0.059	0.115	0.009	0.115	96.74	-0.152	0.122	-0.002	0.122	95.35
MCAR_MNAR	0.058	0.116	0.008	0.116	96.43	-0.13	0.125	0.02	0.126	94.2

Table 6 The estimates of X covariate in scenario 2 for the two outcomes

	True $\beta_{11} = 0.05$					True $\beta_{12} = -0.15$				
Type	Mean	Std	Bias	RMSE	CP	Mean	Std	Bias	RMSE	CP
Complete	0.047	0.107	-0.003	0.107	95.24	-0.15	0.105	0	0.105	94.37
MCAR_MCAR	0.047	0.116	-0.003	0.116	95.24	-0.152	0.119	-0.002	0.119	93.51
MAR_MAR	0.048	0.16	-0.002	0.159	96.54	-0.149	0.132	0.001	0.132	93.94
MNAR_MNAR	0.047	0.12	-0.003	0.12	96.97	-0.139	0.12	0.011	0.121	94.81
COMP_MCAR	0.047	0.106	-0.003	0.106	95.24	-0.151	0.114	-0.001	0.114	93.94
COMP_MAR	0.057	0.155	0.007	0.154	95.24	-0.143	0.138	0.007	0.138	92.64
COMP_MNAR	0.048	0.107	-0.002	0.106	95.24	-0.141	0.12	0.009	0.12	94.37
MCAR_MAR	0.057	0.143	0.007	0.142	94.78	-0.145	0.129	0.005	0.129	93.04
MCAR_MNAR	0.051	0.116	0.001	0.115	94.81	-0.137	0.121	0.013	0.121	93.51

Table 7 The estimates of X covariate in scenario 4 for the two outcomes

	True $\beta_{11} = 0.05$					True $\beta_{12} = -0.15$				
Type	Mean	Std	Bias	RMSE	CP	Mean	Std	Bias	RMSE	CP
Complete	0.054	0.163	0.004	0.163	93.1	-0.18	0.163	-0.03	0.166	91.81
MCAR_MCAR	0.057	0.167	0.007	0.167	93.53	-0.179	0.166	-0.029	0.168	92.24
MAR_MAR	0.134	0.629	0.084	0.63	95.83	-0.114	0.683	0.036	0.68	95.83
MNAR_MNAR	0.045	0.172	-0.005	0.172	93.45	-0.167	0.174	-0.017	0.175	92.14
COMP_MCAR	0.054	0.163	0.004	0.163	93.97	-0.18	0.166	-0.03	0.168	91.81
COMP_MAR	0.052	0.334	0.002	0.332	90.91	-0.146	0.499	0.004	0.496	96.1
COMP_MNAR	0.053	0.164	0.003	0.164	93.48	-0.163	0.162	-0.013	0.162	92.61
MCAR_MAR	0.041	0.418	-0.009	0.416	94.05	-0.18	0.444	-0.03	0.442	94.05
MCAR_MNAR	0.055	0.167	0.005	0.167	93.94	-0.16	0.166	-0.01	0.166	92.64

Citation: Alzahrani H. Missing data analysis for binary multivariate longitudinal data through a simulation study. *Biom Biostat Int J.* 2018;7(2):103–113. DOI: 10.15406/bbij.2018.07.00196

Table 8 The estimates of X covariate for different imputation methods of MCAR_MCAR

	True $\beta_{11} = 0.05$				**True** $\beta_{12} = -0.15$		
	Scen 1	Scen 2	Scen 4		Scen 1	Scen 2	Scen 4
	Mean				Mean		
Complete	0.0529	0.0469	0.0549	Complete	-0.1479	-0.1474	-0.1778
CC	0.0635	0.0396	0.0562	CC	-0.148	-0.1581	-0.1835
LOCF	0.0567	0.0506	0.0572	LOCF	-0.1448	-0.1502	-0.1801
RI	0.026	0.0109	0.0191	RI	-0.1554	-0.1695	-0.1872
MS	0.0571	0.0431	0.0567	MS	-0.1463	-0.1582	-0.181
Incomplete	0.0589	0.0468	0.0566	Incomplete	-0.1443	-0.1518	-0.1791
	std				std		
Complete	0.1116	0.1056	0.1595	Complete	0.102	0.1062	0.1617
CC	0.1413	0.1482	0.2152	CC	0.1396	0.1428	0.2255
LOCF	0.1233	0.126	0.1672	LOCF	0.1238	0.128	0.1654
RI	0.1044	0.1182	0.15	RI	0.0986	0.1096	0.1444
MS	0.1203	0.1196	0.1715	MS	0.1226	0.1234	0.1776
Incomplete	0.1154	0.1161	0.1663	Incomplete	0.1145	0.1188	0.1642
	bias				bias		
Complete	0.0032	-0.0031	0.0062	Complete	0.0018	0.0026	-0.0263
CC	0.0136	-0.0104	0.0072	CC	0.0024	-0.0081	-0.0313
LOCF	0.0073	0.0006	0.0085	LOCF	0.0047	-0.0002	-0.0286
RI	-0.0237	-0.0391	-0.0293	RI	-0.0052	-0.0195	-0.0359
MS	0.0072	-0.0069	0.0077	MS	0.0038	-0.0082	-0.0289
Incomplete	0.0092	-0.0032	0.0079	Incomplete	0.0055	-0.0018	-0.0275
	CP				CP		
Complete	94	95.6	93.6	Complete	96.4	94.4	92
CC	95.11	92.21	95.26	CC	95.11	93.51	91.38
LOCF	94.22	94.81	93.97	LOCF	95.11	93.51	92.67
RI	94.67	93.07	93.53	RI	96	94.81	91.81
MS	92.89	91.77	88.36	MS	92.89	90.91	87.93
Incomplete	96.89	95.24	93.53	Incomplete	95.56	93.51	92.24

In Table 9 the results of handling the mechanism COMP_MCAR for estimated X effects on the log odds of response $P(Y_{ijk}=1)$ .This mechanism is a mixture when the first outcome is complete and the second is MCAR. We found the best method in all the scenarios is MS with good means, less standard deviation and bias for the estimated parameters of the two outcomes unless in the parameter β_{12} in scenario 4. This estimated parameter is already bias in the completed data and MS method have shown a slight reduction in the bias criteria. We found the estimates of the completed outcome were not affected by the MCAR mechanism in the second outcome regardless of the correlation induced between the outcomes or occasions.

In Table 10 where the mechanism is mixed between the MCAR for the first outcomes and MAR for the second outcome. Here it is a good point to testify the stability of the estimated parameters of the first outcome when the parameters of the second outcome has MAR set up pattern for its missing data. We found the MS method has been good for all the scenarios for the two outcomes with good means and less std. This generalization has exception in the second parameter β_{12} in scenario 1 and β_{11} in scenario 4 when the there a slight difference between the incomplete data and MS results. Here we indicated the mean substitution imputation reduce the bias of the MAR missing mechanism in the second outcomes and doing good results to impute the first outcome when its mechanism is MCAR. Finally, it seems the MS method has good means and less standard deviation and less bias in most the results of handling the missing data. In Table 11, we have mixed missing mechanisms for the two outcomes. While the first outcome is complete, the second outcome has MAR assumption for its missing data. Here the goal is testify the precision of the

Citation: Alzahrani H. Missing data analysis for binary multivariate longitudinal data through a simulation study. *Biom Biostat Int J.* 2018;7(2):103–113. DOI: 10.15406/bbij.2018.07.00196

parameters of first outcomes if the they will be affected by the MAR assumption, especially when the two outcomes are correlated. Here the the additional method IPW is added since it is appropriate to reduce the bias in the MAR assumption. Generally, we found the RI method has good means, CP and less std for all the scenarios while the CC has the worst results. Based on the bias results we found, RI and MS and IPW reduced the bias from the completed data in scenarios 2 and 4 where correlation is increased. Generally, we can conclude from model parameters the MS and IPW have some good results to treat the missingness in the mechanism COMP_MAR.[14]

Table 9 The estimates of covariate for different imputation methods of COMP_MCAR

	True $\beta_{11} = 0.05$				**True** $\beta_{12} = -0.15$		
	Scen 1	Scen 2	Scen 4		Scen 1	Scen 2	Scen 4
	Mean				Mean		
Complete	0.0529	0.0469	0.0549	Complete	-0.1479	-0.1474	-0.1778
CC	0.0635	0.0396	0.0562	CC	-0.148	-0.1581	-0.1835
LOCF	0.0554	0.0465	0.0538	LOCF	-0.1444	-0.1514	-0.1802
RI	0.055	0.046	0.0536	RI	-0.1624	-0.1632	-0.1936
MS	0.0553	0.0471	0.053	MS	-0.1496	-0.1497	-0.1794
Incomplete	0.0554	0.0465	0.0538	Incomplete	-0.1444	-0.1514	-0.1802
	std				std		
Complete	0.1116	0.1056	0.1595	Complete	0.102	0.1062	0.1617
CC	0.1413	0.1482	0.2152	CC	0.1396	0.1428	0.2255
LOCF	0.112	0.1063	0.1623	LOCF	0.1145	0.1145	0.1646
RI	0.1118	0.1063	0.1634	RI	0.0998	0.1032	0.1434
MS	0.1124	0.1075	0.1626	MS	0.1213	0.1263	0.1765
Incomplete	0.112	0.1063	0.1623	Incomplete	0.1145	0.1145	0.1646
	bias				bias		
Complete	0.0032	-0.0031	0.0062	Complete	0.0018	0.0026	-0.0263
CC	0.0136	-0.0104	0.0072	CC	0.0024	-0.0081	-0.0313
LOCF	0.0057	-0.0035	0.0051	LOCF	0.0053	-0.0014	-0.0286
RI	0.0053	-0.004	0.0049	RI	-0.013	-0.0132	-0.0421
MS	0.0056	-0.0029	0.0043	MS	0.0004	0.0003	-0.0278
Incomplete	0.0057	-0.0035	0.0051	Incomplete	0.0053	-0.0014	-0.0286
	CP				CP		
Complete	94	95.6	93.6	Complete	96.4	94.4	92
CC	95.11	92.21	95.26	CC	95.11	93.51	91.38
LOCF	94.22	95.24	93.97	LOCF	96	93.94	91.81
RI	94.22	94.81	93.97	RI	96.89	93.07	92.24
MS	94.22	94.37	93.97	MS	93.78	89.61	87.07
Incomplete	94.22	95.24	93.97	Incomplete	96	93.94	91.81

Table 10 The estimates of covariate for different imputation methods of MCAR_MAR

	True $\beta_{11} = 0.05$				**True** $\beta_{12} = -0.15$		
	Scen 1	Scen 2	Scen 4		Scen 1	Scen 2	Scen 4
	Mean				Mean		
Complete	0.0529	0.0469	0.0549	Complete	-0.1479	-0.1474	-0.1778
CC	0.0733	0.111	0.025	CC	-0.1025	-0.1188	-0.0963
LOCF	0.0586	0.0495	0.0562	LOCF	-0.1701	-0.1699	-0.1817
RI	0.0227	0.0152	0.0087	RI	-0.1536	-0.1608	-0.1777
MS	0.0534	0.048	0.0598	MS	-0.1433	-0.15	-0.1421
Incomplete	0.059	0.0571	0.0413	Incomplete	-0.1519	-0.1453	-0.1799
	std				std		
Complete	0.1116	0.1056	0.1595	Complete	0.102	0.1062	0.1617
CC	0.2222	0.2773	0.3348	CC	0.204	0.2185	0.2938
LOCF	0.1141	0.1159	0.1659	LOCF	0.136	0.1342	0.1644
RI	0.1017	0.1091	0.1441	RI	0.0894	0.0954	0.1223
MS	0.1284	0.1264	0.1757	MS	0.152	0.1391	0.174
Incomplete	0.115	0.1426	0.4179	Incomplete	0.1219	0.1291	0.4438
	bias				bias		
Complete	0.0032	-0.0031	0.0062	Complete	0.0018	0.0026	-0.0263
CC	0.0224	0.061	-0.0232	CC	0.0478	0.0312	0.0546
LOCF	0.0089	-0.0006	0.0075	LOCF	-0.0206	-0.0199	-0.0299
RI	-0.0269	-0.0348	-0.0399	RI	-0.0041	-0.0108	-0.0265
MS	0.0038	-0.002	0.0108	MS	0.0066	0	0.0092
Incomplete	0.0093	0.0071	-0.0087	Incomplete	-0.0024	0.0047	-0.0299
	CP				CP		
Complete	94	95.6	93.6	Complete	96.4	94.4	92
CC	93.78	93.51	92.24	CC	95.11	94.37	95.26
LOCF	96.44	95.24	93.97	LOCF	95.11	94.37	93.1
RI	94.22	92.21	93.97	RI	95.56	92.64	93.53
MS	91.11	89.61	88.36	MS	84	88.31	83.19
Incomplete	96.76	94.78	94.05	Incomplete	95.37	93.04	94.05

Table 11 The estimates of covariate for different imputation methods of COMP_MAR

	True $\beta_{12} = -0.15$				**True** $\beta_{12} = -0.15$		
	Scen 1	Scen 2	Scen 4		Scen 1	Scen 2	Scen 4
	Mean				Mean		
Complete	0.0529	0.0469	0.0549	Complete	-0.1479	-0.1474	-0.1778
CC	0.0536	0.12	0.0353	CC	-0.1037	-0.1099	-0.095
LOCF	0.0552	0.0482	0.0537	LOCF	-0.1706	-0.1691	-0.1828
RI	0.055	0.0456	0.0537	RI	-0.1534	-0.1534	-0.1725
MS	0.0551	0.0458	0.0532	MS	-0.1475	-0.1531	-0.1374
IPW	0.0526	0.0447	0.0532	IPW	-0.153	-0.1521	-0.1585

Citation: Alzahrani H. Missing data analysis for binary multivariate longitudinal data through a simulation study. *Biom Biostat Int J.* 2018;7(2):103–113. DOI: 10.15406/bbij.2018.07.00196

Table Continued

	True $\beta_{12} = -0.15$				**True** $\beta_{12} = -0.15$		
Incomplete	0.0551	0.0575	0.0525	Incomplete	-0.1496	-0.1425	-0.1465
	std				std		
Complete	0.1116	0.1056	0.1595	Complete	0.102	0.1062	0.1617
CC	0.1675	0.1922	0.2599	CC	0.1576	0.1507	0.2181
LOCF	0.1114	0.1062	0.1624	LOCF	0.1358	0.1343	0.1644
RI	0.1111	0.1035	0.1637	RI	0.0964	0.09	0.1144
MS	0.1118	0.1116	0.1634	MS	0.1495	0.1397	0.1773
IPW	0.1718	0.1043	0.1181	IPW	0.3062	0.1792	0.2085
Incomplete	0.1112	0.1545	0.334	Incomplete	0.1209	0.1383	0.4995
	bias				bias		
Complete	0.0032	-0.0031	0.0062	Complete	0.0018	0.0026	-0.0263
CC	0.0036	0.07	-0.013	CC	0.0457	0.0401	0.0568
LOCF	0.0055	-0.0018	0.005	LOCF	-0.0212	-0.0191	-0.031
RI	0.0053	-0.0044	0.0051	RI	-0.004	-0.0034	-0.021
MS	0.0054	-0.0042	0.0044	MS	0.0025	-0.0031	0.0138
IPW	-0.0053	0.0032	0.0026	IPW	-0.0021	-0.0085	-0.003
Incomplete	0.0055	0.0075	0.0025	Incomplete	-0.0001	0.0075	0.0035
	CP				CP		
Complete	94	95.6	93.6	Complete	96.4	94.4	92
CC	93.78	94.37	92.24	CC	94.22	95.67	94.4
LOCF	94.22	95.67	93.1	LOCF	94.67	94.37	93.1
RI	94.22	95.67	93.1	RI	96.44	95.24	96.12
MS	93.78	93.51	93.53	MS	84.44	86.58	84.91
IPW	93.98	95.18	95.52	IPW	93.57	93.57	93.72
Incomplete	93.87	95.24	90.91	Incomplete	94.81	92.64	96.1

Conclusion

We presented the analysis of incomplete multivariate binary longitudinal data. The correlation parameters for the complete and incomplete data are estimated. We found the missingness affects the estimated correlation among the occasions more than between the outcomes. Also, we found the estimates of MAR and MNAR of incomplete data are affected by the induced correlation over the occasions and the outcomes. This agrees with results of about the bias of estimates of MAR or MNAR mechanisms for GEE models. After imputing the incomplete data with four missing data handling methods, we conclude many points. The mean substitution based on the multivariate structure mostly has good estimates in the mechanism COMP_MCAR, MCAR_MAR and COMP_MAR. It has been a good remark to find the MS imputation reduced the effects of MAR assumption in the mixed mechanisms and generated mostly less biased results. In the mechanism MCAR_MCAR, the LOCF method has good results. Also, using the weighted GEE for the full and mixed MAR assumptions shows some good results. Generally, we recommend using the mean substitution based on the multivariate structure to impute the missing data. It could be a good future work to use the multiple imputation to handle the missingness in the multivariate structure based on Shelton et al.[7] model.

Acknowledgments

None.

Conflict of interest

Author declares there is no conflict of interest.

References

1. Little RJ, Rubin DB. *Statistical analysis with missing data*. US: John Wiley & Sons; 1987.
2. Myers WR. Handling missing data in clinical trials: an overview. *Drug Information Journal*. 2000;34(2):525–533.
3. Touloumi G, Babiker A, Pocock S. Impact of missing data due to drop-outs on estimators for rates of change in longitudinal studies: a simulation study. *Statistics in medicine*. 2001;20(24):3715–3728.

Citation: Alzahrani H. Missing data analysis for binary multivariate longitudinal data through a simulation study. *Biom Biostat Int J.* 2018;7(2):103–113. DOI: 10.15406/bbij.2018.07.00196

4. Newman DA. Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*. 2003;6(3):328–362.

5. Hening DA. Missing Data Imputation Method Comparison in Ohio University Student Retention Database. USA: Ohio University; 2009.

6. Ali AM, Dawson SJ, Blows FM, et al. Comparison of methods for handling missing data on immuno histochemical markers in survival analysis of breast cancer. *Br J Cancer*. 2011;104(4):693–699.

7. Shelton BJ, Gilbert GH, Liu B, et al. A SAS macro for the analysis of multivariate longitudinal binary outcomes. *Comput Methods Programs Biomed*. 2004;76(2):163–175.

8. Allison PD. *Multiple regression:A primer.* USA: Pine Forge Press; 1999.

9. Alzahrani H. Generating multivariate longitudinal binary random variables for gee models using bridge distribution. *International Journal of Advanced Research*. 2017;5(2):163–175.

10. Hedeker D, Gibbons RD. Longitudinal data analysis, 451 volume. USA: John Wiley & Sons; 2006.

11. Cook RJ, Zeng L,Yi GY. Marginal analysis of incomplete longitudinal binary data: a cautionary note on locf imputation. *Biometrics*. 2004;60(3):820–828.

12. Shao J, Zhong B. Last observation carry–forward and last observation analysis. *Stat Med*. 2003;22(15):2429–2441.

13. Preisser JS, Lohman KK, Rathouz PJ. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Stat Med*. 2002;21(20):3035–3054.

14. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 1995;90(429):122–129.

Citation: Alzahrani H. Missing data analysis for binary multivariate longitudinal data through a simulation study. *Biom Biostat Int J.* 2018;7(2):103–113. DOI: 10.15406/bbij.2018.07.00196