Chelonian Conservation And Biology

Vol. 19 No. 1 (2024) | https://www.acgpublishing.com/ | ISSN - 1071-8443

DOI: doi.org/10.18011/2024.01(1).1994-2005

MODELLING RESPONSE WITH TIME-VARYING COMPLIANCE IN LONGITUDINAL DATA: A SIMULATION STUDY WITH TWO-STAGE FRAMEWORK USING COMPLIANCE REGRESSION RESIDUALS IN CLINICAL TRIALS

Sayyed Sarfaraz*
M.Sc., Department of Analytics, Novartis Healthcare Pvt. Ltd, Telangana, Hyderabad, India
*Corresponding Author Email ID: sarfaraz.sayyed@novartis.com

Ashwini Mathur
PhD, Onesto Consulting, Dublin, Ireland

Asha Kamath
PhD, Department of Data Science, Prasanna School of Public Health, Manipal Academy of Higher Education, Manipal, Karnataka, India

Abstract:

Background: Clinical studies, such as randomized controlled trials, typically measure response and key event incidence throughout the follow-up period. Patients may skip certain assessment visits due to lack of efficacy or safety concerns. Missing data being a common problem in statistical literature, approaches to handle it may still result in biased knowledge discovery. Analysis and interpretation become problematic when missing data percentage is substantial. Additionally, compliance to planned treatment paradigm could also be a problem as patients might not adhere to prescribed treatment regimen. Twin consequences of non-compliance and missing data are rarely addressed simultaneously, even though numerous innovative techniques to handle non-compliance or missing response in randomized trials have been proposed.

Materials and Methods: This article attempted to address the missing response by deploying 2 stage modelling for analysing longitudinal response using time-varying compliance regression residual. Given the variation in longitudinal outcomes, accounting for the dependency between continuous response and treatment compliance can be informative, especially for imputing missing data. EM algorithm is used in this process and compared with/without 2 stage modelling. Simulation study is created with missing response and non-compliance to assess the effectiveness of proposed estimators in scenarios, including both continuous and binary treatment compliance.

Results: Method was applied on simulated data with varying correlation and multiple missing scenarios in both the cases. The results were compared using the absolute bias and mean squared error (MSE).

Conclusions: The MSE was smallest for the proposed method compared to without joint modelling and no imputation analysis, indicating better results with the proposed method.

Keywords: *Missing data; Joint modelling; EM algorithm; non-compliance; 2 stage modelling*

INTRODUCTION:

Whenever response missingness is observed in randomized controlled trials (RCT) with interventional, they are imputed using some conservative approach but doesn't consider the compliance of the treatment. Adherence to study protocol is necessary for analysis of RCT using standard methodologies, which includes obtaining response from all participants, to have unbiased estimates. In an RCT with two arms, non-compliance or non-adherence to the treatment is observed when subjects don't follow treatment regimen as prescribed or discontinue treatment before course completion. Nonresponse happens when required response measurement is missing. If we only include non-missing observations in the analysis, then we typically end up with a smaller sample size. It will affect the power, variability and might produce biased results.

Not all missing values exhibit the same behaviour and mechanism to which they are broadly categorized into 3 categories by Rubin and Little[1] as;

a. "Missing completely at random (MCAR)":

Probability of missing assessment does not depend on the observed values.

b. "Missing at random (MAR)":

Probability of missing assessment depends on the observed values.

c. "Not missing at random (NMAR)":
d. Probability of missing assessment depends on observed as well as unobserved values.

MATERIAL AND METHODS:

Notations and Distribution: Let Y_{ij} , C_{ij} denote the response and compliance from subject 'i' for given time-point 'j', where i = 1, · · ·, n and j = 1, · · ·, t in longitudinal setting. Define the vector of response as $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{it})'$ and the vector of compliance as $C_i = (C_{i1}, C_{i2}, \dots, C_{it})'$. The response data follows a multivariate gaussian distribution. An identity function is used to link the expected continuous response $E(Y_{ij}) = \mu_{ij}$ to the systematic component $X_{ij}'\beta_2$ where X_{ij}' is the matrix of independent variables and β_2 is the vector of corresponding coefficients. The longitudinal compliance variable is assumed to follow longitudinal Beta distribution when compliance is continuous and $0 \leq C_{ij} \leq 1$ or binomial distribution when compliance is binary and $P_c = p(C_{ij} \leq c)$ is linked to a linear function of covariates $X_{ij}'\beta_1$ via a traditional logit function.

In a clinical trial we usually end up having some of the response data missing. This missing data would impact the power of the study and may lead to biased results. Since the response data has missing values, we propose a two-stage modelling[2] framework to address this issue. In this two-stage modelling framework, the first stage involves fitting a longitudinal beta/binomial regression to time-varying treatment compliance, and the second stage uses the residuals from this regression as covariate in a regression model for the longitudinal continuous response.

Stage 1: Longitudinal Beta/Binomial Regression on time-varying treatment compliance: The first stage models the time-varying compliance C_{ij} using beta regression[13,14].

$$Logit(P_c) = X_{ij}'\beta_1 + \varepsilon_{1i} \qquad \text{Equation 1}$$

X_{ij}' represents a vector of covariates that affect C_{ij} such as baseline covariates, time effects, or other subject-specific covariates and $\varepsilon_{1i} \sim N(0, \sigma_1^2)$ is the error term. Once the longitudinal beta/binomial regression is fit, compute the residuals R_{ij} for each time point and subject as the difference between the observed value C_{ij} and the predicted value $\widehat{C_{ij}}$:

$$R_{ij} = C_{ij} - \widehat{C_{ij}}$$

Where $\widehat{C_{ij}}$ is the predicted mean from the beta regression model.

Stage 2: Regression on Longitudinal Continuous Response Data

In the second stage, we model Y_{ij} using the residuals R_{ij} from the first stage, along with any additional covariates X_{ij}' (e.g., time, subject-specific covariates).

The model for Y_{ij} can be formulated as a linear mixed-effects model to account for repeated measurements on the same subjects:

$$Y_{ij} = X_{ij}'\beta_2 + R_{ij}\beta_3 + u_i + \varepsilon_{2i} \quad \text{Equation 2}$$

where:

- R_{ij} are the residuals from the longitudinal beta regression (Stage 1).
- X_{ij}' are additional covariates for the longitudinal continuous response.
- β_2 & β_3 are the vector of corresponding coefficients
- u_i is random effect for subject 'i' to account for within-subject correlation over time.
- $\varepsilon_{2i} \sim N(0, \sigma_2^2)$ is the error term.

Full Likelihood

The full likelihood of this two-stage model can be written as:

$$L(Y, C; \beta_1, \beta_2, \beta_3, u_i, \sigma^2) = \prod_{i=1}^{n} \prod_{j=1}^{t} \left[f_{Beta}\left(C_{ij} \middle| X_{ij}', \beta_1\right) \times f_{Normal}\left(Y_{ij} \middle| X_{ij}', R_{ij}, \beta_2, \beta_3, u_i, \sigma^2\right) \right]$$

where:

- $f_{Beta}\left(C_{ij} \middle| X_{ij}', \beta_1\right)$ is the likelihood of the beta regression (Stage 1).
- $f_{Normal}\left(Y_{ij} \middle| X_{ij}', R_{ij}, \beta_2, \beta_3, u_i, \sigma^2\right)$ is the likelihood of the linear mixed-effects model for the longitudinal continuous response (Stage 2).

The non-response and treatment nonadherence will be first addressed independently using EM algorithm[1,3] and then we further deploy the 2-stage modelling to further refine the estimates of the response.

EM algorithm

The EM algorithm[1,3] is a computational technique in the maximum likelihood estimate (MLE) computation for circumstances when the complexities in computing the MLE are caused by incomplete observation in the data due to MAR, where parameters are separate for observation and mechanism of missing data allowing to ignore the missing data mechanism.

Let the complete response be partitioned between observed response and missing response as

$$\mathrm{Y} = (Y_{obs}, Y_{mis})$$

The log-likelihood for complete set of data is given by,

$$L(\theta) = \log[l(\theta; Y_{obs}, Y_{mis})] = log[f(Y_{obs}, Y_{mis};\ \theta)],$$

The log-likelihood of the marginal or incomplete data is based on y alone and is given by,

$$L_{y_{obs}}(\theta) = \log[l(\theta; Y_{obs})] = log[f(Y_{obs};\ \theta)]$$

We wish to maximize $L_{y_{obs}}(\theta)$ in θ but $L_{y_{obs}}(\theta)$ is typically quite unpleasant:

$$L_{y_{obs}}(\theta) = log \int f(Y_{obs}, Y_{mis};\ \theta) dy_{mis} \quad \text{Equation 1}$$

The EM algorithm is a method that aims at maximizing Equation 3 iteratively while looping between the 2 steps, the E-step that is shown below in Equation 4 and the M-step in the Equation (5) until convergence or a pre-defined threshold met. The E-step uses the current estimated parameter to find expectation for the complete log-likelihood of the data while the M-step makes use of the data updated in E-step to derive the MLE for the parameters.
The expectation step represented here as Q function for the **E-step** is given by,

$$Q\left(\theta/\theta^{(m-1)}\right) = E_{\theta^{(m-1)}}\left(log\frac{f(Y_{obs},Y_{mis};\,\theta)}{f(Y_{obs},Y_{mis};\,\theta^{(m-1)})}|Y_{obs}\right)$$
$$= \int log\frac{f(Y_{obs},Y_{mis};\,\theta)}{f(Y_{obs},Y_{mis};\,\theta^{(m-1)})} f(Y_{mis}|\,Y_{obs};\,\theta^{(m-1)})dy_{mis} \qquad \text{Equation 2}$$

The **M-step** maximizes $Q\left(\theta/\theta^{(m-1)}\right)$ in θ for for fixed $\theta^{(m-1)}$, i.e., it calculates,

$$\theta^{(m)} = arg\ max_{\theta}\ Q\left(\theta/\theta^{(m-1)}\right) \qquad \text{Equation 3}$$

Addressing missingness in compliance
To address missingness in compliance we would partition it as the observed compliance and missing compliance data as

$$C = (C_{obs}, C_{mis})$$
$$f(C_{obs}, C_{mis}/\Psi) = f(C_{mis}/C_{obs}, \Psi_{\text{obs}})f(C_{obs}/\Psi_{\text{mis}}),$$

Where Ψ is the vector of parameters for compliance representing β_1 from Equation 1.

The expectation step represented here as R function for the **E-step** is given by

$$R\left(\Psi/\Psi^{(m-1)}\right) = E_{\Psi^{(m-1)}}\left(log\frac{f(C_{obs},\,C_{mis};\,\Psi)}{f(C_{obs},\,C_{mis};\,\Psi^{(m-1)})}|C_{obs}\right)$$
$$= \int log\frac{f(C_{obs},\,C_{mis};\,\Psi)}{f(C_{obs},\,C_{mis};\,\Psi^{(m-1)})} f(C_{mis}|\,C_{obs};\,\Psi^{(m-1)})dc_{mis} \qquad \text{Equation 4}$$

The **M-step** will maximize the $R\left(\Psi/\Psi^{(m-1)}\right)$ in Ψ for fixed $\Psi^{(m-1)}$, i.e., it calculates,

$$\Psi^{(m)} = arg\ max_{\Psi}\ R\left(\Psi/\Psi^{(m-1)}\right) \qquad \text{Equation 5}$$

Here again we iterate between **E-step** defined in equation 6 and **M-step** as defined in equation 7 to obtain the optimum estimates until convergence or pre-defined threshold met.
The missingness in compliance will be following 'MAR' mechanism. We roughly consider 15% of the compliance data missing. This includes intermittent missing as well as missing due to discontinuation or lost to follow-up. We evaluate 2 cases as below.

1) Longitudinal continuous compliance: Details under **Creating compliance** section.
2) Longitudinal binary compliance: Here the compliance is generated as in equation 8 and after influencing the response with the continuous compliance the compliance data is made binary with 70% or above as compliant or event met (1) and below 70% as non-compliant or event not met (0).

Longitudinal compliance was used in modelling case 1 whereas for case 2 longitudinal binary compliance was converted to univariate proportion between (0, 1) depending on number of visits where subject met the compliant criteria.

Simulation with non-response and non-compliance

We started with sample size of 554 subjects (227 subjects per arm) for obtaining treatment difference of around 1 unit at last visit to be simulated from 4 time point Multivariate Gaussian distribution using sample size derivation for longitudinal study[4,5]. The correlation structure used was Autoregressive of order 1 (AR1) with three values for ρ (0.4, 0.6, 0.8) the most reflected in RCT, but other correlation structures can be used as well. Artificial missingness was introduced in the data to observe behaviour of the methods. Missingness introduced in simulated data was roughly 5%, 10%, 15% and 20% using MAR approach.

Simulation and analysis were repeated 1000 times to reflect robustness of the methods.

Introducing artificial missingness in the data

For simulation purpose we considered making the in data using MAR mechanism.

For "MAR" mechanism, first we calculate weighted sum scores. Weighted sum scores are a linear combination of the variables. We will use the method as explained in Schouten and Vink, 2018[6] to make data missing with MAR mechanism of missingness.

The process requires a complete data with n subjects having responses at m time-points. The process results in multiple sub datasets with either complete or incomplete subsets. All these sub-datasets are then combined to obtain a version from original dataset that contains pre-specified missingness in data.

The process starts with determining required missing data patterns. For example, lost to follow-up case from visit 4 in study with 5 time-points, subjects will not be missing for first 3 time-points and will be missing for last 2 time-points.

First random division of original complete dataset into k sub-datasets takes place on basis of k missing data patterns. Size for these sub-datasets may not be same. For instance, if we assign the frequency value of one-third for one of the missing data patterns and two-third for another missing data pattern, this then will result in one-third cases becoming part of sub-dataset 1 and two-third cases becoming part of sub-dataset 2. These subjects in sub-dataset 1 will become the candidates for first missing data pattern. The frequency values should sum-up to 1 so that each participant falls in one of the sub-datasets. Up till here we are just preparing the subset of candidates falling in each of the k sub-datasets pertaining to a particular missing pattern. We will then calculate weighted sum score (SSw) for each subject.

The sum of the weighted score of subject 'i' is calculated using the following equation:

$$SSw_i = wt_1.Y_{i1} + wt_2.Y_{i2} + \cdots + wt_t.Y_{it} ,$$

where $\{Y_{i1}, Y_{i2}, \ldots, Y_{it}\}$ are responses of subject 'i' at time-points 1,2, …, t and $\{wt_1, wt_2, \ldots, wt_t\}$ are corresponding weights that are pre-specified. SSw will be largely influenced by variables with higher weights as compared to variables with low weights. Weights can be positive or negative based on relative importance of the variables.

Since under MAR mechanism the missing depends on observed values, weights will be zero corresponding to values that will be made incomplete. This will differ for each pattern. Figure 1 shows schematic overview about amputation procedure for introducing missingness in data.

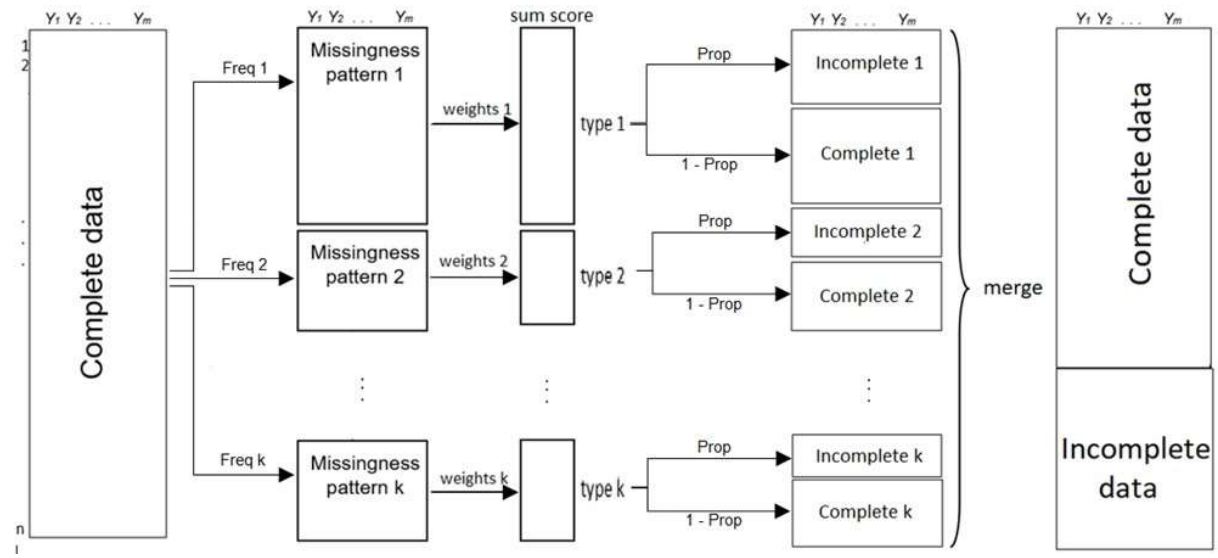

Figure 1 Schematic overview of procedure for amputation in the multivariate/longitudinal data

For given variable based on SSw each subject will be assigned the probability of becoming non-responder. To distribute these probabilities, left skewed logistic distribution function is applied on SSw (Figure 2). For different patterns of missing data differently skewed logistic distribution function can be used.

Logistic distribution function used for obtaining the probability of missingness

Logit is the log of odds given by

$$logit(p) = \log\left(\frac{p}{1-p}\right), with\ 0 < p < 1$$

logistic function is obtained by equating logit to possible regressors set x and solving for p.

$$p = \frac{exp\ (x)}{1 + exp\ (x)}$$

We used left skewed logistic function for creating missingness in complete data. This resembles more with real scenario where patient's response is more likely to be missing as patient progresses in the trial rather at the start. So low probability of missing at initial time-point and high probability of having response missing at study end.

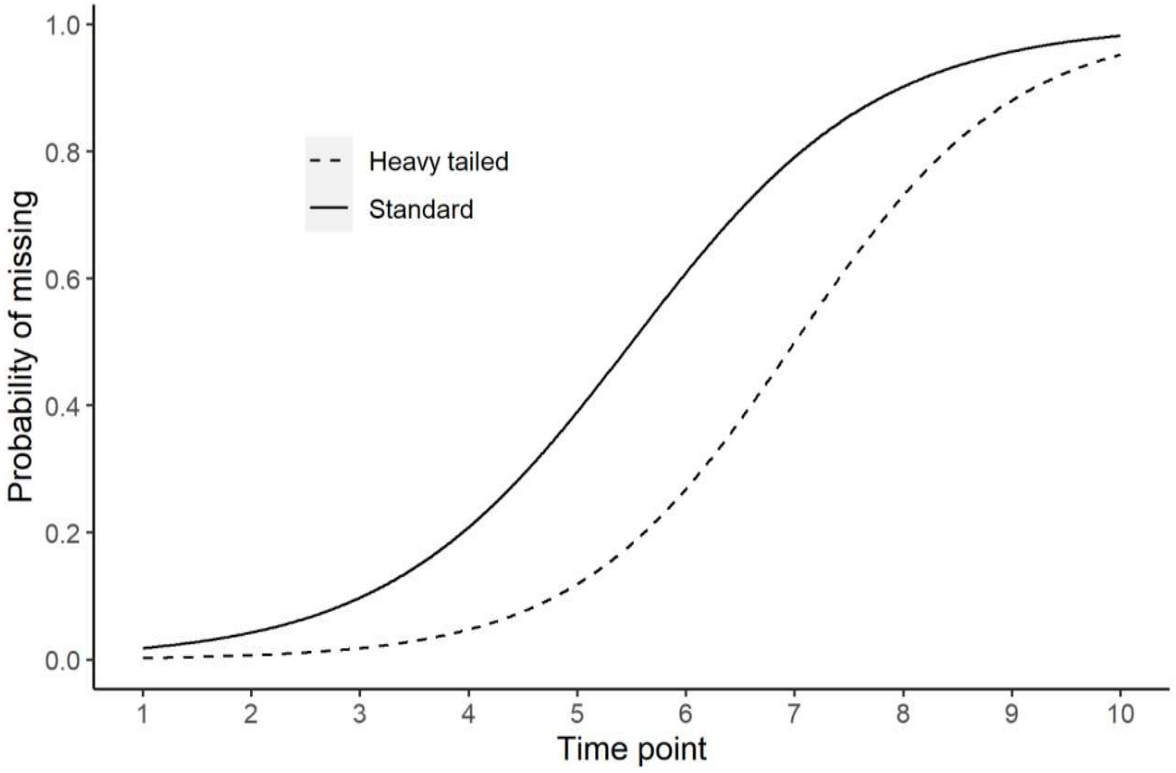

Figure 2 Standard and heavy tailed logistic distribution functions

Creating Compliance

Compliance is constructed as function of median compliance, 10th percentile compliance and compliance covariance matrix[7]. Noncompliance is assumed to alter responses by "regressing" them towards baseline values.

True compliance ($C^{(g)}$) for group 'g' will be discrete if we count actual doses taken during study period but if we consider the proportions such as 28/28, 26/28, etc., then the value $C^{(g)}$ can be seen as continuous approximations.

True compliance might be actually reflecting variety of different behaviors including dose timing, food compliance and concomitant medications taken. In all this type of cases using continuous distribution for $C^{(g)}$ has better justification.

We create compliance as continuous assessment [0, 1] scale, where 1 would denote perfect or 100% compliance and 0 denotes 0% or no compliance.

Our initial inputs are **10th percentile of compliance and median compliance** and using median as 0.9, and 10th percentile as 0.3, we say that 50% of the patients are 90% or better compliant, and nearly 90% of the patients are 30% or better compliant.

Since there is subject-specific effect as well as carryover effect for compliance data, it makes more sense to use CS (compound symmetry) + AR1 (Auto regressive of order 1) model.

We start by generating Z_{ij} ('i' denotes patient, 'j' denotes time-point) from multivariate normal distribution with covariance structure as CS + AR(1) and then using pre-specified median compliance and 10th percentile of compliance to generate proportions by using the transformation of normal probability shown in equation 8 below.

$$C_{ij}^{(g)} = \phi\left(a^{(g)} + b^{(g)} Z_{ij}\right) \qquad \text{Equation 6}$$

$$Where, \quad a^{(g)} = \phi^{-1}\left(C_{0.5}^{(g)}\right) \;\&\; b^{(g)} = \frac{\phi^{-1}\left(C_{0.1}^{(g)}\right) - \phi^{-1}\left(C_{0.5}^{(g)}\right)}{\phi^{-1}\left(C_{0.1}^{(g)}\right)}$$

$C_{ij}^{(g)}$denotes compliance rate for i[th] subject at j[th] time-point for group g, ϕ denotes Z-distribution value and ϕ^{-1} denotes area under Z-distribution curve for given point.
Since noncompliance is expected to influence treatment response, below equation is used to transform responses towards baseline values.

$$y_{ij}^{(g)\prime} \leftarrow y_{ib}^{(g)} + C_{ij}^{(g)}(y_{ij}^{(g)} - y_{ib}^{(g)})$$

where $C_{ij}^{(g)}$ denotes "compliance rate" and will lie between 0 and 1.
$y_{ij}^{(g)}$denotes observation from i[th] subject in g[th] group at j[th] time-point and $y_{ib}^{(g)}$denotes observation from i[th] subject in g[th] group at baseline.

The model and the likelihood

For modelling the response, a linear mixed effects model for repeated measures was fitted[5]. Kenward-roger method was used to estimate denominator degrees of freedom. The aim was to estimate treatment effect against comparator and hence 2 treatments parallel arm design setting was used. The imputation was done separately in each arm.
For equation 1 compliance was fitted against Drug, time, Drug x time interaction and baseline compliance as covariates and for equation 2 the covariates for response included treatment given, time of response assessment, treatment-by-time interaction, residual from the equation 1 and baseline response.
We assumed AR(1) type of correlation structure for creating responses in simulated dataset but while analyzing no assumption on variance-covariance was made and hence unstructured variance-covariance was used.
The simulation was repeated 1000 times and each of these datasets were analyzed separately and results from each dataset were pooled using Rubin's rule[8] as follows:

$$Pooled\ estimate\ \bar{\theta} = \frac{\sum_{i=1}^{m}\theta_i}{\mathrm{m}}\ ,Where\ \theta_i\ is\ estimate\ from\ individual\ simulation$$

$$Pooled\ standard\ error = \sqrt{V_w + \left(1 + \frac{1}{m}\right)V_B},$$

$$Where\ V_w\ is\ mean\ of\ squared\ individual\ simulation\ SE's\ and\ V_B = \frac{\sum_{i=1}^{m}(\theta_i - \bar{\theta})^2}{\mathrm{m} - 1}$$

Heatmap shown in Figure 3 gives the visualization about deployed missing pattern.

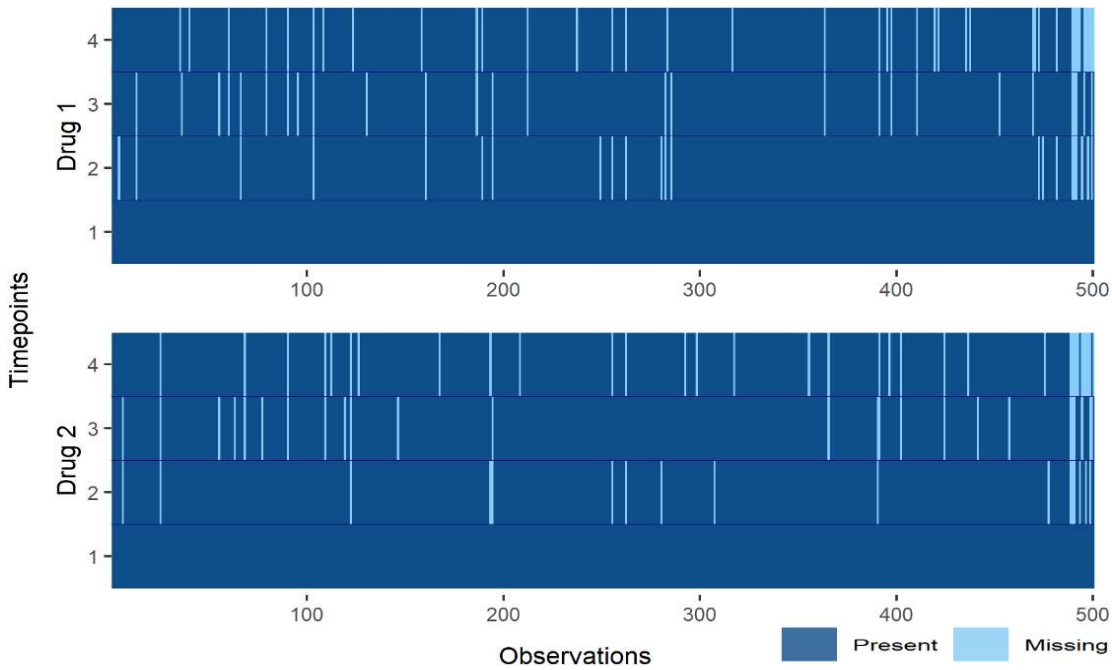

Figure 3 Missing pattern and the Missingness in the rows in one of the simulated data

RESULTS:

Table 1: Simulation results with both response and compliance as continuous outcome

		2 Stage modelling		Response imputation		Without imputation	
Correlation	**Missing%**	**Estimate (SE)**	**Absolute Bias**	**Estimate (SE)**	**Absolute Bias**	**Estimate (SE)**	**Absolute Bias**
0.4	10	0.9468(0.2602)	0.0012	0.9314(0.26)	0.0143	0.9299(0.2684)	0.0158
	20	0.9464(0.2606)	0.0008	0.9314(0.2605)	0.0143	0.9269(0.2783)	0.0187
	30	0.9466(0.261)	0.001	0.9317(0.2609)	0.014	0.9232(0.2895)	0.0225
	40	0.9463(0.2614)	0.0006	0.9307(0.2613)	0.0149	0.9179(0.3018)	0.0278
0.6	10	0.9187(0.2341)	0.0017	0.9427(0.2334)	0.0223	0.952(0.2411)	0.0316
	20	0.9183(0.2351)	0.0021	0.9389(0.2345)	0.0185	0.9578(0.2511)	0.0374
	30	0.9213(0.2357)	0.0008	0.9388(0.2353)	0.0183	0.9621(0.2622)	0.0416
	40	0.9243(0.236)	0.0039	0.9393(0.2357)	0.0189	0.9648(0.2741)	0.0444
0.8	10	0.9425(0.2644)	0.0031	0.9007(0.2598)	0.0388	0.8942(0.268)	0.0453

	20	0.939(0.2648)	0.0005	0.8991(0.2604)	0.0404	0.8886(0.2781)	0.0508
	30	0.9423(0.2653)	0.0028	0.9046(0.2611)	0.0349	0.8849(0.2896)	0.0545
	40	0.9451(0.2659)	0.0057	0.9092(0.262)	0.0303	0.8801(0.3022)	0.0593

Table 2: Simulation results with continuous response and binary compliance outcome

		2 Stage modelling		**Response imputation**		**Without imputation**	
Correlation	**Missing%**	**Estimate (SE)**	**Absolute Bias**	**Estimate (SE)**	**Absolute Bias**	**Estimate (SE)**	**Absolute Bias**
0.4	10	0.9451(0.2614)	0.0005	0.9451(0.2604)	0.0005	0.9435(0.2688)	0.0021
	20	0.9466(0.2618)	0.0009	0.9466(0.2609)	0.0009	0.9415(0.2788)	0.0041
	30	0.9464(0.2621)	0.0008	0.9464(0.2612)	0.0008	0.9391(0.29)	0.0066
	40	0.9453(0.2626)	0.0003	0.9453(0.2617)	0.0003	0.936(0.3023)	0.0096
0.6	10	0.9231(0.2342)	0.0026	0.9231(0.2338)	0.0026	0.931(0.2416)	0.0105
	20	0.9214(0.2351)	0.0009	0.9214(0.2348)	0.0009	0.9383(0.2516)	0.0179
	30	0.9283(0.2357)	0.0079	0.9283(0.2355)	0.0079	0.9444(0.2626)	0.024
	40	0.9319(0.2361)	0.0114	0.9319(0.2359)	0.0114	0.9487(0.2745)	0.0282
0.8	10	0.9395(0.2657)	0.0001	0.9395(0.2614)	0.0001	0.9365(0.2698)	0.003
	20	0.9378(0.2661)	0.0017	0.9378(0.2621)	0.0017	0.9327(0.28)	0.0068
	30	0.9365(0.2666)	0.0029	0.9365(0.2627)	0.0029	0.9299(0.2915)	0.0096
	40	0.94(0.2673)	0.0005	0.94(0.2635)	0.0005	0.9266(0.3042)	0.0128

Pooled estimates from simulated studies are presented in Table 1 & Table 2 after applying the proposed method and with standard method of imputation without joint modelling. Additionally, the tables also present outcomes from analysis of observed cases without imputation which in current simulation environment can be a useful benchmark to check against induced bias due to missing data. We have presented results with different percentage of missingness in dataset ranging from 10% to 40% having moderate to high correlation (0.4, 0.6 and 0.8). All estimates are generated assuming 'MAR' mechanism. Table 1 provides results when responses are continuous longitudinal as well as compliance is continuous longitudinal whereas Table 2 provides results

when responses are continuous longitudinal, but compliance is longitudinal binary. The MSE is comprised of both the components, variance and the squared bias. The aggregated standard error was calculated using Rubin's rule.

Under all correlation types with continuous response and continuous compliance, our proposed method produces estimates with lower MSE and more than 90% lower bias when compared to analysis without imputation or analysis with imputation and without 2 stage modelling (Table 1).

With respect to data with continuous response and binary compliance we did not notice any kind of improvement as compared to Imputation without 2 stage modelling. Though they both were better than no imputation method and reduced the bias more than 70% as compared to analysis without imputation. No improvement over Imputation without 2 stage modelling might be due to high threshold set for making the compliance data binary(>70%) (Table 2).

DISCUSSION:

Performance of EM algorithm along with joint modelling provides better estimates as compared to EM algorithm alone and further estimation is much better as compared to no imputation with respect to reducing the bias as seen in Table 1 and Table 2 for with differently correlated data as well as for different percentage of missingness in the data.

We found estimates from proposed techniques are nearer to population parameter than estimates from non-imputed data. Some interesting conclusions have arisen out of the analysis. As results improve using treatment compliance effect into the model, it indicates compliance can play a significant role in imputing missing observations for obtaining better estimates.

It's quite evident from results that presence of missing data, has high bias, and one must use appropriate missing data handling technique to reduce bias. We in our research have only considered case of continuous response, further exploration to see performance of method in other types of response data might be required.

Our proposed technique is to use EM algorithm and then improvise it with 2 stage modelling, another alternative to this would be to use Multiple imputation (MI) and then improvise it with 2 stage modelling. In both cases we found that estimates from our proposed methods are closer to assumed population parameter than estimates from non-imputed data. But to reach convergence, repetition required for EM algorithm in cases with large percentage of missingness is more than that of MI. This could be overcome by replacing M-step of EM algorithm with one step Newton–Raphson[9] to speed the convergence.

Also, with respect to computational time, standard EM algorithm turns out to be quite faster as compared to MI. Time taken by MI method is around fivefold more than EM algorithm when tried for an imputation using MI with 10 copies of datasets at each repetition.

There are certain criticisms for MI based on computing and analysis time. For e.g., it is costlier to analyze 10 sets of data as compared to one analysis. Similar critical assessment made by Fay[10] was that the use of MI should be in large and public-use datasets where individual who is imputing data and one who is analyzing should be separate which is also addressed by Meng[11]. Rubin[12] also made a note that model used in generating MI datasets should have all variables that would likely be used in subsequent analyses.

Since computations are done on simulated data, no specific data are associated with this article. All computations are done using 'R-4.2.1' software, codes can be obtained upon request to authors.

REFERENCES:

1. Little RJA, Rubin DB. Statistical Analysis with Missing Data, New York, Wiley, 2002; Second Edition.
2. Angrist JD, Imbens GW. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. J Am Stat Assoc. 1995; 90(430):431-442.
3. Little RJA. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. J Am Stat Assoc. 1995; 90(431):1112-1121.
4. Sayyed S, Mathur A, Kamath A. Sample size variation in single-time post-dose assessment vs multi-time post-dose assessment. F1000Res. 2022; 11:1550.
5. Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of Longitudinal Data, New York, Oxford University Press, 2002; Second Edition
6. Schouten RM, Lugtig P, Vink G. Generating missing values for simulation purposes: a multivariate amputation procedure. J Stat Comput Simul. 2018; 88(15):2909-2930.
7. Kimko HC, Duffull SB. Simulation for Designing Clinical Trials. CRC Press, 2019; First Edition.
8. Rubin DB. Multiple Imputation for Nonresponse in Surveys, Wiley, 1987; First Edition.
9. Herzet C, Wautelet X, Ramon V, Vandendorpe L. Iterative Synchronization: EM Algorithm Versus Newton-Raphson Method. IEEE International Conference on Acoustics Speed and Signal Processing Proceedings. 2006; pp. IV-IV.
10. Fay RE. When are Inferences from Multiple Imputation Valid? Proc. Survey Res Meth Sec. Am Stat Assoc; 1986; 227-232.
11. Fay RE. Multiple-Imputation Inferences with Uncongenial Sources of Input: Comment. Statistical Science. 1994; 9(4):558-560.
12. Rubin DB. Multiple Imputation After 18+ Years. J Am Stat Assoc. 1996; 91(434):473-489.
13. Ferrari, S.L.P., & Cribari-Neto, F. Beta regression for modeling rates and proportions. Journal of Applied Statistics. 2004; 31(7): 799-815.
14. Smithson, M., & Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods. 2006: 11(1): 54–71.