A Simulation Study to Compare Missing Data Imputation Methods For Binary Outcome Variables

Fang Liu[*] Jingjing Chen[†]

Abstract
The analysis of longitudinal data in clinical trials presents a challenge as there are often missing data points. When binary outcomes variables are involved, the missing data imputation methods may become complicated. A simulation study illustrates how Generalized Linear Mixed Model (GLMM), Inverse Probability Weighted (IPW) Generalized Estimation Equation (GEE) method, multiple imputation and doubly robust method work in practice, especially for binary outcome variables in terms of efficiency and accuracy, with MAR assumption.

Key Words: missing at random, multiple imputation, inverse probability weighting, IPW GEE, longitudinal data

1. Introduction

Longitudinal data plays an important role in clinical trials, in which same individuals are measured repeatedly on some important variables according to a pre-determined schedule. Despite efforts to minimize missing data in the design and conduct of clinical trials, missing data are still encountered often due to a variety of reasons and a major source of missing data occurs when subjects drop out of study.

Existing guidelines for the design and conduct of clinical trials, and the analysis of the resulting data, provide only limited advice on how to handle missing data (National Research Council, 2010). Thus approaches to the analysis of data with missing values are still challenging, especially when binary response variables are involved. In this paper, we focus on approaches applied to monotone missing pattern, which means missingess occurs only through dropout. We also limited our discussion to missing at random (MAR), where missingness depends on observed data but not on unoberseved data.

The most popular approach to analyze longitudinal data with binary responses is Generalized Estimation Equation (GEE) method (Liang and Zeger, 1986). As Liang and Zeger pointed out, inference with the GEE method is valid only under the stronger assumption that the data are missing completely random (MCAR), where missingness is independent of both unobserved and observed data. Therefore, Robins et al. (1995) proposed a class of weighted estimating equation in which observations have weights inversely proportional to the probability of being observed. This weighted estimating equation approach, which has been called the inverse probability weighted (IPW) GEE, is valid under MAR (Preisser et. al., 2002). However, IPW GEE yields bias estimate if the model for probability of missing (missingness model) is not correctly specified. Doubly robust estimation methods were developed to remedy this weakness, see Van der Laan and Robins (2003); Bang and Robins (2005); Seaman and Copas (2009). Doubly robust methods involve both a missingness model and an imputation model for the expectation of each missing observation, and are consistent when either is correct. Thus, it offers more protection against model misspecification than IPW GEE.

[*] Merck & Co. Inc, 351 North Sumneytown Pike, Upper Gwynedd, PA, 19454

[†] MedImmune, LLC, One MedImmune Way, Gaithersburg, MD, 20878

An alternative approach to analysis of incomplete longitudinal data is use of Generalized linear mixed model (GLMM)(Breslow and Clayton, 1993), which attracted considerable attention after SAS's GLIMMIX procedure became available. GLMM is an extension to the linear mixed model, which allows response variables with different distributions, such as binary responses. It can incorporate both fixed and random effects in the model. By modeling the individual subject variables as a random effect, it allows for the accommodation of multiple missing data points.

Another well-known method to handle missing data is multiple imputation (MI), which was developed by Rubin (1987). The key idea of this procedure is to fist impute missing data several times, then analyze the resulting complete data with standard methods for complete data, such as GEE. These analyses generate a set of results that are afterwards appropriately combined to provide a single estimate of the parameter of interest, together with standard errors that reflect the uncertainty inherent in the imputation of the missing data (Fitzmaurice et al. 2011). The basic form of MI requires MAR, though versions under MNAR have been proposed.

This paper is intended to evaluate the performance of Generalized Linear Mixed Model, multiple imputation, IPW GEE and Doubly Robust estimation methods, when being applied to longitudinal data in clinical trial with missing binary responses, to assess safety or efficacy between two drug products , i.e., test (T) and reference (R).

We organize this paper as follows. In Section 2, an overview of approaches for analyzing longitudinal data with missing binary response variables is given with focus on generalized linear mixed models, multiple imputation, IPW GEE and doubly robust estimation methods. Section 3 reports simulation results of comparing these four methods in terms of bias and standard error of the estimates, type-I error control and power. We end with a discussion of the relative advantages of different approaches in Section 4.

1.1 Background

We consider a longitudinal study with T visits for each individuals. Let $Y_i = (Y_{i1}, ..., Y_{iT})'$ be the binary response vector for individual i at visit t ($i = 1, 2, ..., N$; $t = 1, 2, ..., T$). Likewise, let $X_i = (X_{i1}', ..., X_{iT}')'$ represents the corresponding $p \times n$ covariates matrix, where X_{it} is a $p \times 1$ vector of covariates associated with Y_{it}.

We define $R_{it} = 1$ if Y_{it} is observed, and 0 otherwise. Note that under monotone missing pattern, if $R_{it} = 0$, then $(R_{i,t+1}, ..., R_{iT}) = (0, ..., 0)$.

Let $\mu_i = E(Y_i|X_i)$. We assume that Y_i given X_i, follows a generalized linear model with mean

$$\mu_i = g^{-1}(X_i'\beta), \tag{1}$$

where g is the link function and β is a $p \times 1$ vector of unknown parameters of interest. For example, for a binary outcome following the logistic regression model, $g(p) = log\frac{p}{1-p}$.

1.2 Inverse Probability Weighted Generalized Estimating Equations Approach

The IPW-GEE weights each individual's contribution to the GEE by the inverse probability of being observed up to a certain timepoint. Using a similar notation to Fitzmaurice at al. (2011), the IPW-GEE estimate of the parameters β are the solution from the following weighted estimating equations:

$$\sum_{i=1}^{N} \frac{\delta\mu_i}{\delta\beta'} V_i^{-1} W_i (Y_i - \mu_i) = 0, \tag{2}$$

where V_i is a $T \times T$ working covariance matrix of Y_i and W_i is a $T \times T$ diagonal matrix of the occasion-specific weights, $W_i = diag(R_{i1}w_{i1}, ..., R_{iT}w_{iT})$, which will be described below (see equation (3)).. Note that the weights is given by w_{it} for observed responses and 0 for unobserved responses.

In practice, the IPW GEE estimate can be obtained from the following two-step procedure:

Step 1: fit a model for $\pi_{it} = Pr\{R_{it} = 1|R_{i1} = ... = R_{i,t-1} = 1, X_{i1}, ..., X_{it}, Y_{i1}, ..., Y_{i,t-1}\}$, for example, using logistic regression (the missingness model). The weights at observed visits is computed as the inverse of the cumulative conditional probabilities of remaining in the study:

$$w_{it} = (\pi_{i1} \times ... \times \pi_{it})^{-1}. \tag{3}$$

Step 2: apply GEE model using PROC Genmod with weight statement (the final analysis model).

It was proved that IPW GEE gives unbiased estimates under MAR when both the missingness model and the final analysis model are correctly specified. If the missingness model is misspecified, the resulting estimates may be biased.

1.3 Doubly Robust Estimation Method

There are many possible ways of obtaining a doubly robust estimation in the literature. Here we use the doubly robust method developed by Hernandez et. al. (2014), as it can be implemented using standard SAS procedures. Hernandez et. al. (2014) extended the idea from Vansteelandt et al. (2010) to longitudinal data, since the doubly robust estimate method proposed by Vansteelandt et al. (2010) can be applied to non-longitudinal data only.

Before considering a doubly robust estimate, Hernandez et. al. (2014) divided all the covariates X into two sets: The first set Z contains covariates whose coefficients we wish to estimate (e.g., treatment, time); these variables are to be included in the final analysis model. The second set F explains the relationship between the response variable Y and the missingness in the data. These variables should be included in the missingness and imputation models but not the final analysis model. Following Vansteelandt et al. (2010), Hernandez et. al. (2014) developed a three-stage approach to obtain doubly robust estimates as below:

Step 1: fit a logistic regression model for the probability of being observed as a function of F and Z (Missingness model). Let π_{it} denote the fitted probability and calculate the w_{it} as in (3) .

Step 2: fit a generalized linear model to the response Y using all covariates F and Z for the data where Y is observed with Weight statement (imputation model).

Step 3: replace the values of the response Y with the fitted values $m^*(F, Z)$ as calculated in step 2. Fit another generalized linear model for all subjects (both fully observed and those that had a missing outcome) to the new response $m^*(F, Z)$ using only the covariates Z (final analysis model).

Since the response values from the final analysis generalized linear model (step 3) are the predicted values from the previous weighted GLM (step 2), the true variance is underestimated. Thus, the standard error and p-value of the estimated treatment effect are obtained using the bootstrap procedure. The resulting coefficient estimates will be consistent if either the missingness model in step 1 or the imputation model in step 2 is correctly specified, provided that the final analysis model in step 3 is correctly specified.

1.4 Generalized Linear Mixed Model

A generalized linear mixed model is given by:

$$g[E(Y|X)] = X\beta + Z\gamma + \epsilon, \tag{4}$$

where γ is an unknown vector of random parameters with design matrix Z and $\epsilon \sim N(0, \sigma^2)$. Unlike the linear mixed model, estimation of the GLMM using maximum likelihood requires an iterative process. One method is called restricted pseudo-likelihood, which estimates a pseudo-response by maximizing the residual log pseudo-likelihood with a first-order Taylor series expansion around the solutions of the best linear unbiased predictors of the random effects (Wolfinger and OConnell, 1993; SAS Institute Inc., 2011). Due to the pseudo-likelihood estimation of the marginal population-average fixed effects, the logistic regression GLMM assumes data are MCAR. Although GEE and GLMM both require MCAR, if data is MAR and the correlation structure of the repeated measurements are correctly specified, the logistic regression GLMM generally provides a less bias estimate of the fixed effects than GEE (Zeger, Liang and Albert,1988).

The optimization method in SAS GLIMMIX is specified with NLOPTIONS TECHNIQUE=NRRIDG. PROC GLIMMIX does not have a REPEATED statement; however, covariance structures are modeled with the RANDOM statement, using the RSIDE option (Davis 2014). When the model does not converge with an unstructured covariance pattern, other covariance structure might be considered and a supportive model should be run by adding EMPIRICAL option to obtain the empirical sandwich estimator.

1.5 Multiple Imputation

Several imputation methods for binary response variables are provided in SAS procedure PROC MI, such as logistic regression and Markov Chain Monte Carlo (MCMC). Since MCMC is mainly used for arbitrary missing data, we introduce multiple imputation with the logistic regression method. Multiple imputation can be carried out with a three-step procedure as well:

Step 1: impute the missing response using the imputation model for a number of times (M) and produces M imputed datasets. With monotone missing pattern where Y_{i1} is fully observed, missing values in the second response Y_{i2} can be imputed by fitting an appropriate model (e.g., a logistic regression model) to predict Y_{i2} from Y_{i1} and X_i. Then missing values in the third response Y_{i3} can be imputed based on an appropriate model to predict Y_{i3} from Y_{i1}, X_i and both observed and imputed values of Y_{i2}. Imputation of remaining missing values can continue in a similar way until all of the missing values have been filled in (Fitzmaurice et. al., 2011). In the next section of simulation study, 20 imputed datasets were generated.

Step 2: analyze each of the M imputed datasets separately with the GEE model (final analysis model).

Step 3: pool the analysis results obtained from step 2 from the M imputed datasets into one single reference. The combined point estimate and variance of the parameter of interest β are given by

$$\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{M} \qquad \text{and} \qquad V = W + (\frac{M+1}{M})B, \tag{5}$$

where

$$W \;=\; \sum_{m=1}^{M}\frac{W^m}{M} \qquad \text{and} \qquad M = \sum_{m=1}^{M}\frac{(\hat{\beta}^M - \bar{\hat{\beta}})(\hat{\beta}^M - \bar{\hat{\beta}})'}{M-1}, \tag{6}$$

Table 1: Mean Response of Active and Control

Case	Treatment	E(Y)
1	Active	(0.40, 0.45, 0.50, 0.55, 0.60)
	Control	(0.40, 0.45, 0.50, 0.55, 0.60)
2	Active	(0.40, 0.41, 0.42, 0.44, 0.45)
	Control	(0.40, 0.45, 0.50, 0.55, 0.60)

with W denoting the average *within* imputation variance and B the *between* imputation variance (Rubin, 1987; Birhanu et. al., 2011).

For the multiple imputation method to be valid, all covariates that are needed to ensure that the response is MAR should be included in the imputation model. Omitting one variable could introduce bias in parameter estimates and result in poor estimates. For final analysis model fitted to imputed datasets, it can include only covariates of interest.

2. Simulation Study

2.1 Data Generation

We constructed a simulation experiment to compare the above four methods under different conditions pertaining to significance of parameter of interest and the amount of missingness. In the simulated data, 400 individuals received one of the two treatments with ratio $1:1$, e.g., active and control. For each individual, a vector of correlated binary responses from five visits $Y_i = (Y_{i1}, ..., Y_{i5})'$ were generated with common correlation $\rho = 0.5$ using an algorithm developed by Park et. al. (1996). The responses between individuals were assumed independent. Two cases in Table 1 were considered for the mean binary responses $E(Y)$.

Case 1: no difference between active and control. Empirical type-I error was examined as the proportion of not detecting the difference at visit 5 from 500 simulations.

Case 2: significant difference of mean binary response at visit 5 between active and control. Empirical power was calculated as the proportion of detecting the difference at visit 5 among 500 simulations.

The missingness process was assumed to be MAR, and the probability of being observed at visit t for each treatment was modeled by a logistic regression of the form

$$logit(P(R_{it} = 1)) = \alpha + \beta * Y_{i,t-1} + \gamma * Base\_stat_i, t = 2, .., 5, \tag{7}$$

where $Base\_stat$ is a categorical variable with random values from $(1, 2, 3)$, which is independent to the response variable. Different sets of α, β and γ values for each treatment were determined to yield four scenarios of dropout rate in table 2.

2.2 Simulation Results

Simulation results are summarized in Tables 3 to 6. Tables 3 and 4 presents the results of cases 1 and 2, where everything is correctly specified in all methods. The type-I error for doubly robust estimate seems to be inflated in certain scenarios. For example, the empirical type-I error rate is 0.068 when dropout rate is within the range of $15-20\%$ and $25-30\%$ for active and control treatment respectively. Therefore, the doubly robust estimate is

Table 2: Drop Out Rate

Scenario	Active	Control
1	15-20%	15-20%
2	15-20%	25-30%
3	25-30%	15-20%
4	25-30%	25-30%

Table 3: Simulation Result of Case 1 with Correct Models

	Empirical Type I error				95% CI Coverage				Bias of Log(OR) Estimate			
Scenario	GLMM	MI	IPW	DR	GLMM	MI	IPW	DR	GLMM	MI	IPW	DR
1	0.044	0.052	0.044	0.058	0.96	0.94	0.95	0.94	0.00	0.00	-0.03	-0.03
2	0.048	0.046	0.054	0.068	0.95	0.95	0.94	0.93	-0.01	-0.01	-0.09	-0.09
3	0.052	0.050	0.048	0.054	0.95	0.95	0.95	0.95	0.00	0.00	0.04	0.04
4	0.050	0.054	0.048	0.044	0.95	0.95	0.94	0.96	-0.01	0.00	-0.02	-0.02

not considered further. Though some type-I error rates from GLMM, multiple imputation and IPW GEE methods are higher than 0.05 (maximum 0.054) due to the variation from simulation, we consider these three methods are able to control type-I error at level 0.05. In addition, IPW GEE method generates larger bias than GLMM and multiple imputation in both cases, though it seems to be more powerful.

Tables 5 and 6 report results when some models are misspecified, in the sense that variable $base\_stat$ is omitted from the missingness model in IPW GEE, imputation model in multiple imputation and GLMM. The results from GLMM appears still reasonable. The performance of IPW GEE becomes unstable with much less power in scenario 2 but more power in other scenarios in case 2. Moreover, it yields relative lower confidence interval coverage and much larger bias in both cases. Multiple imputation is consistently more conservative in both cases. Though it produces minor bias in case 1, the observed bias in case 2 is even larger than IPW GEE in two scenarios.

3. Discussion

Through our simulation study, GLMM seems to perform very well in all cases and is recommended as primary analysis in clinical trial with missing binary response. However, cautions should be taken when using SAS PROC GLIMMIX since the results could be quite different with different options in PROC GLIMMIX and type-I error could be inflat-

Table 4: Simulation Result of Case 2 with Correct Models

	Empirical Power				95% CI Coverage				Bias of Log(OR) Estimate			
Scenario	GLMM	MI	IPW	DR	GLMM	MI	IPW	DR	GLMM	MI	IPW	DR
1	0.78	0.80	0.82	0.80	0.96	0.96	0.95	0.94	-0.01	-0.01	-0.08	-0.05
2	0.78	0.78	0.76	0.72	0.96	0.94	0.95	0.95	0.00	0.00	0.00	0.02
3	0.81	0.81	0.88	0.88	0.96	0.95	0.93	0.92	-0.01	-0.01	-0.12	-0.12
4	0.82	0.80	0.84	0.84	0.96	0.96	0.94	0.95	0.00	0.00	-0.05	-0.04

Table 5: Simulation Result of Case 1 with Incorrect Models

	Empirical Type I error			95% CI Coverage			Bias of Log(OR) Estimate		
Scenario	GLMM	MI	IPW	GLMM	MI	IPW	GLMM	MI	IPW
1	0.042	0.024	0.044	0.96	0.98	0.96	0.00	-0.01	-0.04
2	0.044	0.028	0.110	0.96	0.97	0.89	0.00	-0.01	-0.16
3	0.050	0.028	0.066	0.95	0.97	0.93	0.00	0.00	0.11
4	0.054	0.030	0.056	0.95	0.97	0.94	0.00	-0.01	-0.02

Table 6: Simulation Result of Case 2 with Incorrect Models

	Empirical Power			95% CI Coverage			Bias of Log(OR) Estimate		
Scenario	GLMM	MI	IPW	GLMM	MI	IPW	GLMM	MI	IPW
1	0.79	0.58	0.85	0.96	0.96	0.92	-0.01	0.13	-0.12
2	0.78	0.64	0.61	0.96	0.95	0.94	0.00	0.11	0.08
3	0.82	0.65	0.97	0.96	0.96	0.82	-0.01	0.11	-0.27
4	0.81	0.72	0.86	0.96	0.96	0.94	0.00	0.09	-0.06

ed with incorrect covariance structure.

Both IPW GEE and multiple imputation could be good candidates for sensitivity analysis, if done carefully. IPW GEE with a correctly specified missing model is generally less efficient than multiple imputation with a correctly specified imputation model (Seaman and White, 2011). However, when both models are misspecified, multiple imputation becomes too conservative and may produce more bias than IPW GEE in some cases.

Though doubly robust estimation methods are promising theoretically, some of these methods are difficult to implement in SAS or other standard software and it seems that there is a risk of Type-I error inflation through our simulation study. Therefore, Type-I error rates should be examined with simulation before considering doubly robust estimation methods.

REFERENCES

Bang, H., and Robins, J.M. (2005), “Doubly robust estimation in missing data and causal inference models,” *Biometrics*, 61, 962972.

Birhanu,T., Molenberghs, G., Sotto, C., and Kenward, M.G. (2011), “Doubly robust and multiple-imputation-based generalized estimating equations,” *Journal of Biopharmaceutical Statistics*, 21, 202–225.

Breslow, N.E., and Clayton, D.G. (1993), “Approximate Inference in Generalized Linear Mixed Models”. *Journal of the American Statistical Association*, 88 (421), 9–25.

Davis, S. (2014), *Clinical Trials with Missing Data: A Guide for Practitioners*,New York, John Wiley & Sons, Chapter 3.

Fitzmaurice, G.M., Laird, N.M., and Ware, J.H., (2011), *Applied Longitudinal Analysis, Second Edition*, New York, John Wiley & Sons.

Hernandez, B., Lipkovich, I., and O’Kelly, M., and Ratitch, B. (2014), *Clinical Trials with Missing Data: A Guide for Practitioners*,New York, John Wiley & Sons, Chapter 8.

National Research Council (2010), *The Prevention and Treatment of Missing Data in Clinical Trials, Panel on Handling Missing Data in Clinical Trials*, Washington, D.C., The National Academies Press.

Park, C.G., Park, T., and Shin, D.W. (1996), “A simple method for generating correlated binary variates,” *The American Statistician*, 50,306–310.

Preisser, J.S., Lohman, K.K., and Rathouz, P. J. (2002), “Performance of Weighted Estimating Equations for Longitudinal Binary Data with Drop-outs Missing at Random,” *Statistics In Medicine*, 21, 3035–3054.

Robins, J. M., Rotnitzky, A., and Zhao, L.P. (1995), “Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data”. *Journal of the American Statistical Association*, 90, 106–121.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York, John Wiley & Sons.

SAS Institute Inc. (2011), “SAS/STAT 9.3 Users Guide: The GLIMMIX Procedure (Chapter),” SAS Institute Inc., Cary, NC.

Seaman, S., and Copas, A. (2009), “Doubly robust generalized estimating equations for longitudinal data,” *Statistics in Medicine*, 28, 937–955.

Seaman, S.R., and White, I.R. (2011), “Review of inverse probability weighting for dealing with missing data,” *Statistical Methods in Medical Research*, 22, 278–295.

Van Der Laan, M.J., and Robins, J.M. (2003), “Unified Methods for Censored Longitudinal Data and Causality,” New York, Springer.

Vansteelandt, S., Carpenter, J., Kenward, M.G. (2010), “Analysis of incomplete data using inverse probability weighting and doubly robust estimators,” *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6, 3748.

Wolfinger R, and OConnell M (1993), “Generalized linear mixed models: a pseudo-likelihood approach,” *Journal of Statistical Computation and Simulation*, 48, 233243.

Zeger, S.L., Liang, K.Y. and Albert P.S. (1988), Models for Longitudinal Data: A Generalized Estimating Equation Approach, *Biometrics*, 44, 1049–1060.