To adjust or not to adjust for baseline when analyzing repeated binary responses? The case of complete data when treatment comparison at study end is of interest

DOI: 10.1002/pst.1682
Handle: http://hdl.handle.net/1942/18958

Adjusting for baseline on the analysis of repeated binary responses with missing data

Honghua Jiang[1], Pandurang M Kulkarni[1], Craig H Mallinckrodt[1], Linda Shurzinske[1], Geert Molenberghs[2] and Ilya Lipkovich[3]

1. Lilly Research Labs. Eli Lilly and Co. Indianapolis, IN. 46285.
2. I-BioStat, Hasselt University, Diepenbeek, Belgium, and I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium
3. Quintiles

Abstract

Covariate adjusted and unadjusted implementations of the following methods were compared in analyzing incomplete repeated binary data when the outcome at the study endpoint is of interest: logistic regression with the last observation carried forward (LOCF), generalized estimating equations (GEE), weighted GEE (WGEE), generalized linear mixed model (GLMM), and multiple imputation with analyses via GEE (MI). Incomplete data mimicking several clinical trial scenarios were generated using missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) mechanisms. Across the various analytic methods and scenarios covariate adjusted analyses generally yielded larger treatment effect estimates and larger standard errors compared with their unadjusted counterpart. The net result of these factors was increased power from the covariate adjusted analyses without increasing type I error rates. Although all methods were biased in at least some of the MNAR scenarios the type I error rates from LOCF exceeded 30% whereas the highest rate from any other method in any scenario was 10%. LOCF also yielded biased results in MCAR and MAR data whereas the other methods where not biased or had smaller biases than LOCF. These results support longitudinal modeling of repeated binary data over LOCF logistic regression of the study endpoint only. These results also support covariate adjustment for baseline severity in these longitudinal models.

Key words: Missing Data, binary data, covariate adjustment

1. **Introduction**

Binary outcomes derived from underlying continuous measures are commonly evaluated in clinical trials. For example, in diabetes clinical trials glycated haemoglobin (HbA1c) is a continuous measure that reflects average plasma glucose for several months (Sacks 2002). It is commonly used as the primary efficacy outcome measure. However, a clinically meaningful outcome is whether the endpoint HbA1c reaches the target of <7.0% (ADA 2013). Therefore, comparing the probabilities of reaching HbA1c target of <7.0% between treatments based on a binary outcome is an important objective in diabetes trials.

In longitudinal clinical trials patients are treated over a period of time and are evaluated at multiple time points. Usually, the primary efficacy evaluation is based on the measurement at the last scheduled time point. However, patients may withdraw before completing the study and the measurement at the last scheduled time point will be missing. Missing data occur commonly in longitudinal studies for various reasons, including lack of efficacy, safety, re-location, etc. An historically common approach to handle missing data was to impute the missing observations with the last available observation of the patient, i.e., the last observation carried forward (LOCF) method. However, LOCF requires restrictive assumptions that are unlikely to hold in practice and this approach is generally not acceptable (NRC, 2010)

Other analytic approaches for repeated binary data that do not require imputation include generalized estimating equations (GEE) (Liang and Zeger, 1986; Zeger and Liang, 1986). Under the missing completely at random mechanism (MCAR) (Rubin, 1976), GEE provides unbiased and consistent parameter estimates even when the working correlation matrix is mis-specified. However, under the

missing at random mechanism (MAR) (Rubin, 1976) parameter estimates based on GEE can be biased. A weighted generalized estimating equations approach (Robins et al. 1995; and Fitzmaurice et al. 1995) extends conventional GEE and provides consistent parameter estimates under MAR when the dropout model is correctly specified. In this approach, an individual's contribution is weighted by the inverse probability of dropout at the given time.

Likelihood-based generalized linear mixed model (GLMM) analyses of the available cases have also been widely used for the analysis of repeated binary data, generally under the assumption of MAR although some approximations to direct-likelihood require the more stringent MCAR assumtion (Wolfinger and O'Connell, 1993).

Multiple imputation approaches (Rubin, 1987; Schafer 1999; Shieh 2003; Li et al., 2006) are commonly applied to continuous incomplete longitudinal data and can therefore be used to impute the continuous outcome from which the binary responses are derived without some of the restrictive assumptions that limit LOCF. After imputation, the resulting complete data sets can be analyzed with either GEE or likelihood-based methods.

Lipkovich et al. (2005) compared the performance of MI followed by GEE analysis with GLMM and GEE analyses of available cases in estimating treatment differences for binary outcomes derived from underlying continuous responses. The MI-based approach performed better than GEE and GLMM in terms of precision, power, and type I error rate under MAR. However, under the missing not at random (MNAR) mechanism, all three methods yielded biased results. Liu and Zhan (2011) also conducted simulations to compare GLMM, GEE, and several MI approaches for the analysis of repeated binary responses with missing data in evaluating the

treatment effect at study endpoint. Results indicated that GLMM performed better than GEE and MI approaches in terms of controlling type I error rate under MAR.

In many clinical settings the baseline severity is linked to the probability of achieving a target level of symptom severity. For example, in diabetes baseline HbA1c levels influence whether the patient can achieve the HbA1c target of <7.0%. Therefore, adjusting for baseline severity may improve analytic performance in these situations. However, little research has been done to evaluate the effect of adjusting for initial disease severity in the analysis of repeated binary data with missing values. The present research evaluates the performance of the covariate adjusted and unadjusted analyses of repeated binary outcomes in terms of type I error rate, power, precision, and bias across several common statistical approaches. Focus is on evaluating the treatment effect at the study endpoint.

2. **Statistical Methods**

For the analysis of only the outcome at study endpoint, not repeated measures, logistic regression with LOCF was used. For repeated binary outcome analyses GEE, WGEE, GLMM, and MI approaches were used and the study endpoint contrast was derived from the repeated measures analysis. Details of each analysis are described below.

Logistic regression:

For the analysis of a single binary outcome at study endpoint, let Y_i=1 represent that the *i*th patient achieved HbA1c target of <7.0% at the study endpoint, and Y_i=0 otherwise. Let $\boldsymbol{\beta}$ be a vector of regression coefficients; $\boldsymbol{X}_i$ be a vector of covariates such as treatment indicator and baseline HbA1c; and $p_i = \text{pr}(Y_i=1|\, \boldsymbol{X}_i, \beta)=E\,(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta})$. Then the logistic regression model can be expressed as:

$$\text{logit}\,(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X}_i'\,\boldsymbol{\beta} \qquad (1)$$

With LOCF there is no missing data and the key assumption is that patients' observations would not have changed had they stayed in the trial. Logistic regression was implemented in the present study is SAS PROC Logistic (SAS, 2008).

Generalized Estimating Equations (GEE)

For the analysis of repeated binary outcome, let Y_{it}=1 represent that the *i*th patient achieved HbA1c target of <7.0% at time *t*, and Y_{it} =0 otherwise; $\boldsymbol{X}_{it}$ be the vector of covariates for the *i*th patient at time *t*; let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}\,, \ldots\,, Y_{iT}\,)'$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}\,, \boldsymbol{X}_{i2}\,,\ldots, \boldsymbol{X}_{iT}\,)$, and *T* is the number of scheduled study visits at which the data are collected. Let $\boldsymbol{\beta}$ be the vector of regression coefficients; and $\boldsymbol{p}_i\,(\boldsymbol{\beta}) = \text{pr}(\boldsymbol{Y}_i=1|\, \boldsymbol{X}_i, \boldsymbol{\beta})=E\,(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{\beta})$. Then, the GEE proposed by Liang and Zeger (1986) and Zeger and Liang (1986) takes the form

$$\text{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{D}'_i\, \boldsymbol{V}_i^{-1}\big(\boldsymbol{Y}_i - \boldsymbol{p}_i(\boldsymbol{\beta})\big) = \boldsymbol{0} \qquad (2)$$

where $\boldsymbol{D_i} = \partial \boldsymbol{p_i}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and $\boldsymbol{V_i}$ is a "working" covariance matrix. SAS PROC GENMOD (SAS, 2008) was used to implement GEE in the present study. The key missing data assumptions for GEE is that the missing data arise from a MCAR mechanism (Molenberghs and Kenward, 2007).

Weighted Generalized Estimating Equations (WGEE)

The WGEE method proposed by Robins et al. (1995) and Fitzmaurice et al. (1995) is less restrictive than standard GEE in that the key missing data assumption is that the missing data arise from an MAR mechanism - given that the probabilities of dropout for each subject are correctly specified (Molenberghs and Kenward, 2007). In the approach proposed by Robins et al (1995) equation (2) is modified as:

$$\mathrm{U}(\boldsymbol{\beta}) = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \frac{\mathbf{1}}{\boldsymbol{\pi_i}} \boldsymbol{D'_i} \boldsymbol{V_i^{-1}} \left(\boldsymbol{Y_i} - \boldsymbol{p_i}(\boldsymbol{\beta})\right) = \mathbf{0} \qquad (3)$$

where $\boldsymbol{\pi_i}$ is the vector of the probability of dropout for the *i*th subject. So an individual's contribution is weighted by the inverse probability of dropout at each assessment time; that is, a different weighting is applied to each visit. In contrast, the approach of Fitzmaurice et al. (1995) uses a single weight for each patient that is applied to all assessment times. In the present study WGEE was implemented using proc logistic (SAS, 2008) to estimate the dropout probabilities and PROC Genmod (SAS, 2008) was used to incorporate the weightings and conduct analysis.

Generalized Linear Mixed Model (GLMM)

The GLMM extends the generalized linear model by incorporating normally distributed random parameter for individual subjects (Breslow and Clayton, 1993). The fixed-effect inference is conditional on random parameters and has a subject-specific interpretation. The form of GLMM based on the logit link function to fit the response probability, p_{it}, at time point t is as

$$\text{logit}\,(p_{it}) = \boldsymbol{X}'_{\boldsymbol{it}}\,\boldsymbol{\beta} + \boldsymbol{Z}'_{\boldsymbol{it}}\,\boldsymbol{b}_{\boldsymbol{i}} \qquad (4)$$

where $\boldsymbol{\beta}$ is the vector of the fixed-effect parameters, $\boldsymbol{b}_{\boldsymbol{i}}$ is the vector of random subject parameters and $\boldsymbol{b}_{\boldsymbol{i}} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{V})$, and $\boldsymbol{X}_{\boldsymbol{it}}$ and $\boldsymbol{Z}_{\boldsymbol{it}}$ and the vectors of known covariates. In the present study GLMM was implemented using PROC GLIMMIX in SAS (SAS, 2008).

Multiple Imputation (MI):

MI is an extension of single imputation where the missing data are imputed several times, say m times. Then each of the m completed data sets is analyzed with the standard methods and the results of the m analyses are combined according to Rubin's rule [Rubin, 1987]. The following process was implemented for the multiple imputation method.

1. Bayesian regression which included earlier values as predictors was used to impute the missing data with a separate predictive distribution for each treatment group [Rubin, 1987]. The missing continuous HbA1c values were imputed first with SAS MI procedure (SAS, 2008). Next, the continuous outcomes were dichotomized into binary response data according to whether they were <7.0% or not. We used m=30 times to generate m complete data sets.

2. The GEE model as previously described was used to analyze each of the 30 imputed complete data sets for the repeated binary response, and 30 sets of parameter estimates were obtained.

3. The 30 estimates and associated standard errors were combined into the final estimates with SAS PROC MIANALYZE (SAS, 2008) according to Rubin's rule.

3. **Simulation**

3.1 Simulation setting

The continuous HbA1c values (%) at each visit were simulated based on a multivariate normal distribution with mean profiles (Table 1) and variance-covariance matrix with elements $\sigma_{i,j}=\rho\ \sigma_i\ \sigma_j$, where ρ was the correlation between the repeated outcomes. The binary outcome was constructed based on whether the HbA1c value was <7.0% or not. Data were simulated based on inputs obtained from actual diabetes clinical trials. A compound-symmetry correlation matrix was used, with $\rho = 0.5$, and the variance (σ_i^2) increasing over time from visit 1 (at baseline) to visit 4 (1.0, 1.0, 1.2, 1.4). The sample size of 50 and 200 per treatment group were used to mimic phase 2 and phase 3 trial settings, respectively.

Incomplete data sets were then generated from the complete data sets using 3 rates of missing data and 3 missing data mechanisms. Rates of missing data were either 45% in both treatments groups or 25% one treatment group and 45% of in the other. Missingness

mechanisms were either MCAR, MAR, or MNAR. In MCAR the outcomes did not differ for patients that completed compared with those who dropped out. In MAR the probability a value (y_i) being missing depended on the observed value at the previous visit (y_{i-1}), expressed as logit (p (y_i missing | y_{i-1})= $a + b * y_{i-1}$. The value b =1.5 was used in the dropout model and the value a was chosen for each treatment group to achieve the desired rates of missing data. In MNAR the probability of a value (y_i) being missing depended on the value itself (y_i), expressed as

logit (p (y_i missing | y_i)= $a + b * y_i$. The value b =0.4 was used in the dropout model and the value a was chosen for each treatment group to achieve the desired rates of missing data. For simplicity, only monotone missingness was considered. For each scenario 2000 data sets were simulated.

Performance of different analysis methods was evaluated based on bias ($\hat{\beta}$ -β_T) in scenarios when there was no difference between treatments at endpoint, relative bias ($\frac{\hat{\beta} - \beta_T}{\beta_T} \times 100$) in scenarios where treatments did differ at endpoint. Methods were also compared based on 95% confidence interval (CI) coverage (using normal theory approximation), standard errors (SE) (average of SEs from the 2000 simulations), and type I error rate for scenarios with no difference between treatments and power for scenarios where treatments differed. $\hat{\beta}$ is the estimate of log odds ratio for unadjusted or adjusted analysis. β_T is the "true" log odds ratio for the unadjusted or adjusted analysis which is based on estimate from the values obtained by averaging results from the corresponding complete data sets.

Table 1. Mean treatment profiles in simulation model

Hypothesis		visit1 HbA1c (%)	visit2 HbA1c (%)	visit3 HbA1c (%)	visit4 HbA1c (%)
No treatment effect	Treatment	8.5	7.6	7.3	7.0
	Comparator	8.5	7.6	7.3	7.0
Moderate treatment effect	Treatment	8.5	7.6	7.3	7.0
	Comparator	8.5	7.8	7.5	7.2

3.2. **Simulation Results**

Table 2 shows the simulation results under MCAR. LOCF appreciably inflated type I error when rates of missing data differed between treatments. All analyses yielded the anticipated nominal type I error rate when rates of missing data were equal for the two treatments, except for slight inflations with Unadj.logit and Adj.GEE with the small sample size. MI analyses yielded the lowest type I error rate (<3%).

All analyses yielded relatively unbiased estimates with the absolute bias <0.05 for the no treatment effect case, and relative bias <11% for the moderate treatment effect case, except for the LOCF, which had biases 7-18 times greater than other methods for the no treatment effect case with unequal missing proportions, and relative bias up to 62% for the moderate treatment effect case. All analyses produced CIs with coverage close to their nominal level except for LOCF analysis which was associated with coverage lower than nominal level, with the MI analysis exceeding the nominal level. For the moderate treatment effect case, adjusted analyses

yielded greater treatment effect estimates, but also larger SEs compared to the unadjusted analyses, resulting in increased power with the large sample size. However, slightly lower power of the adjusted analyses was observed with the small sample size.

Simulation results under MAR are displayed in Table 3. Unadj.GEE, Unadj.GLMM, WGEE, and LOCF methods inflated type I error with unequal missing data rates. All analyses preserved type I error rate with the equal missing proportion case except for WGEE and Adj.GEE with the small sample size. MI analyses had the lowest type I error rate, <4%. All analyses yielded relatively unbiased estimates with the absolute bias <0.09 for the no treatment effect case, except for LOCF and WGEE, and relatively unbiased estimates for moderate treatment effect case with equal missing data rates (relative bias<=8%). However, for the moderate treatment effect case with unequal missing data rates LOCF and WGEE yielded significantly biased estimates with relative bias up to 114%. In general, adjusted analyses had less bias than their unadjusted counterparts. All analyses produced CIs with coverage close to the nominal level except for LOCF and WGEE, which yielded coverage lower than nominal level, while MI analysis exceeded their nominal level. For the moderate treatment effect case, adjusted analyses yielded greater treatment effect estimates and SEs compared to the unadjusted analyses, but resulted in increased power in general.

Simulation results under MNAR are reported in Table 4. All analyses inflated type I error rate except for MI. For the moderate treatment effect case with equal missing data rates all analyses produced relatively unbiased estimates (relative bias <=10%). However, for the moderate treatment effect case with unequal missing data rates all analyses yielded significantly biased estimates with relative bias up to 57%.

Table 2: Simulation results under MCAR

		No treatment effect								Moderate treatment effect													
		(45%, 45%)*				(25%, 45%)*				(45%, 45%)*						(25%, 45%)*				(45%, 25%)*			
N		Bias	SE	Cov (%)	Rej. (%)	Bias	SE	Cov (%)	Rej. (%)	Est	Rel Bias (%)	SE	Cov (%)	Rej. (%)	Est	Rel Bias (%)	SE	Cov (%)	Rej. (%)	Est	Rel Bias (%)	SE	Cov (%)	Rej. (%)
50	Unad.logit	-0.02	0.41	94	**6.1**	0.15	0.41	92	**7.6**	0.29	12	0.42	95	10.6	0.13	-50	0.41	94	6.6	0.42	62	0.42	93	18.1
	Adj.logit	-0.02	0.45	95	4.8	0.18	0.45	93	**7.0**	0.35	8	0.46	95	11.2	0.16	-51	0.45	94	5.6	0.50	57	0.46	93	18.9
	Unad.GEE	-0.02	0.53	95	5.5	0.02	0.50	95	5.3	0.28	9	0.54	95	8.9	0.26	2	0.50	95	7.3	0.28	7	0.51	94	9.3
	Adj.GEE	-0.01	0.60	94	**6.2**	0.01	0.57	95	5.3	0.34	8	0.61	95	8.8	0.32	1	0.57	94	8.3	0.33	3	0.57	94	9.1
	Unadj. GLMM	-0.02	0.53	95	5.2	0.01	0.50	95	5.1	0.28	8	0.54	95	8.0	0.27	2	0.50	95	6.9	0.27	5	0.51	94	8.7
	Adj. GLMM	-0.01	0.61	94	5.6	0.01	0.57	95	5.0	0.35	9	0.62	95	8.5	0.33	3	0.58	95	8.0	0.33	3	0.58	94	8.6
	Unadj.MI	-0.04	0.50	97	2.7	0.01	0.48	98	2.3	0.29	11	0.50	98	5.2	0.28	6	0.48	97	5.2	0.28	10	0.48	96	6.9
	Adj. MI	-0.04	0.57	98	2.2	0.00	0.54	97	2.6	0.35	9	0.58	98	4.5	0.33	4	0.55	98	5.2	0.34	7	0.55	97	6.7
	Unadj. WGEE	-0.01	0.56	95	5.0	0.02	0.52	95	4.7	0.28	7	0.56	96	7.5	0.26	0	0.52	96	6.6	0.28	8	0.52	95	8.0
	Adj.WGEE	-0.02	0.62	95	5.1	0.01	0.58	95	5.1	0.35	8	0.63	95	7.7	0.32	0	0.58	95	7.9	0.33	4	0.58	95	8.7
200	Unadj.logit	0.01	0.20	95	5.1	0.15	0.20	89	**11.2**	0.28	8	0.21	95	26.4	0.13	-51	0.20	90	9.6	0.42	61	0.21	89	53.0
	Adj.logit	0.01	0.22	95	4.8	0.18	0.22	88	**11.8**	0.33	3	0.22	95	30.9	0.15	-54	0.22	87	10.6	0.49	54	0.22	89	60.3
	Unadj.GEE	0.01	0.26	94	5.6	0.01	0.25	95	4.6	0.28	6	0.26	95	17.2	0.27	3	0.25	95	19.2	0.27	3	0.25	96	18.8
	Adj.GEE	0.00	0.30	95	5.4	0.01	0.28	96	4.3	0.33	3	0.30	96	19.9	0.32	0	0.28	94	20.8	0.32	0	0.28	95	20.1
	Unadj. GLMM	0.01	0.26	94	5.6	0.01	0.25	96	4.4	0.27	5	0.26	95	17.0	0.27	3	0.25	95	19.4	0.27	3	0.25	96	18.5
	Adj. GLMM	0.00	0.30	95	5.3	0.01	0.28	96	4.3	0.33	4	0.30	95	19.7	0.32	0	0.28	94	20.4	0.32	0	0.28	95	20.0
	Unadj.MI	0.00	0.25	97	2.7	0.02	0.24	98	2.4	0.27	5	0.25	98	15.9	0.25	-4	0.24	97	15.6	0.28	6	0.24	98	18.1
	Adj. MI	0.00	0.28	98	2.3	0.02	0.27	98	2.2	0.33	2	0.28	98	17.6	0.30	-7	0.27	97	16.3	0.33	3	0.27	97	19.0
	Unadj. WGEE	0.01	0.27	95	5.2	0.01	0.25	96	3.9	0.27	5	0.27	95	15.8	0.26	0	0.26	95	17.4	0.27	6	0.26	96	17.8
	Adj.WGEE	0.00	0.30	96	4.5	0.02	0.28	97	3.5	0.33	4	0.30	96	19.1	0.31	-2	0.28	95	19.1	0.33	2	0.28	96	19.6

* Percentage of missing data for comparator and treatment, respectively; Cov: coverage of 95% CI; Rej: rejection rate; Rel Bias: relative Bias; Est: - log odds ratio; Boldface font indicates that the type I error rate is beyond 2 SEs of simulations.

Table 3: Simulation results under MAR

		No treatment effect										Moderate treatment effect												
		(45%, 45%)*				(25%, 45%)*						(45%, 45%)*				(25%, 45%)*					(45%, 25%)*			
N		Bias	SE	Cov (%)	Rej %	Bias	SE	Cov (%)	Rej. (%)	Est	Rel Bias (%)	SE	Cov (%)	Rej. (%)	Est	Rel Bias (%)	SE	Cov (%)	Rej. (%)	Est	Rel Bias (%)	SE	Cov (%)	Rej. (%)
50	Unadjlogit	0.01	0.42	95	5.1	0.32	0.41	88	**12.4**	0.25	-5	0.43	96	8.1	-0.03	-111	0.42	90	5.2	0.53	105	0.42	90	23.9
	Adj.logit	0.00	0.47	95	5.3	0.38	0.46	87	**13.0**	0.31	-4	0.48	95	9.5	-0.03	-108	0.47	88	4.5	0.66	105	0.47	89	27.3
	Unadj.GEE	0.01	0.54	95	5.4	-0.05	0.50	94	**6.3**	0.25	-5	0.54	94	8.0	0.31	20	0.49	94	9.8	0.19	-26	0.50	94	6.4
	Adj.GEE	0.01	0.61	94	**6.1**	-0.04	0.56	94	5.9	0.31	-3	0.61	94	8.9	0.36	13	0.56	94	10.8	0.27	-15	0.56	94	8.4
	Unad. GLMM	0.01	0.54	94	5.6	-0.09	0.50	93	**6.1**	0.25	-3	0.54	95	7.7	0.35	33	0.50	94	10.2	0.16	-40	0.50	93	5.8
	Adj. GLMM	0.00	0.62	94	5.8	-0.06	0.57	94	5.6	0.32	0	0.61	94	8.4	0.38	20	0.57	94	10.4	0.26	-18	0.57	94	7.8
	Unadj.MI	-0.01	0.54	96	3.5	-0.02	0.50	97	3.2	0.27	3	0.53	97	5.1	0.29	7	0.49	96	6.7	0.25	-9	0.49	97	5.8
	Adj. MI	-0.02	0.62	97	2.8	-0.03	0.57	97	2.9	0.33	2	0.61	97	5.2	0.35	9	0.56	97	6.7	0.30	-5	0.56	97	6.2
	Unadj.WGEE	0.01	0.70	91	**9.1**	-0.22	0.62	91	**9.1**	0.28	6	0.68	92	9.7	0.48	79	0.60	91	14.4	0.05	-83	0.61	90	7.9
	Adj.WGEE	0.00	0.74	91	**9.1**	-0.11	0.66	93	**7.1**	0.34	6	0.72	92	11.0	0.45	39	0.65	92	12.5	0.22	-30	0.65	92	8.3
200	Unadj.logit	0.01	0.21	94	5.9	0.31	0.21	68	**32.0**	0.24	-7	0.21	95	21.0	-0.04	-115	0.21	69	5.8	0.51	98	0.21	78	70.3
	Adj.logit	0.00	0.23	95	5.5	0.37	0.23	63	**37.5**	0.29	-8	0.23	95	24.6	-0.04	-113	0.23	63	5.7	0.62	94	0.23	76	77.9
	Unadj.GEE	0.00	0.26	95	5.4	-0.07	0.24	93	**6.8**	0.24	-8	0.26	95	15.8	0.31	21	0.24	93	24.5	0.17	-33	0.25	94	10.3
	Adj.GEE	0.00	0.30	94	**6.1**	-0.05	0.28	94	5.8	0.30	-6	0.30	95	18.1	0.35	11	0.27	94	25.2	0.25	-21	0.28	94	15.4
	Unadj. GLMM	0.00	0.26	95	5.5	-0.09	0.25	93	**7.3**	0.24	-7	0.26	95	15.7	0.34	31	0.24	93	28.3	0.15	-42	0.25	93	8.7
	Adj. GLMM	0.00	0.30	94	5.7	-0.06	0.28	94	5.9	0.30	-5	0.30	95	17.7	0.37	15	0.28	94	26.3	0.24	-24	0.28	94	14.1
	Unadj.MI	0.00	0.26	96	3.9	0.02	0.24	96	3.5	0.27	-1	0.26	97	15.2	0.24	-10	0.24	96	14.5	0.27	-1	0.24	97	17.5
	Adj. MI	-0.01	0.30	97	3.2	0.02	0.28	96	3.8	0.32	1	0.30	97	16.4	0.30	-8	0.27	97	16.7	0.32	1	0.27	97	19.9
	UnadjWGEE	0.00	0.37	95	5.2	-0.21	0.31	89	**10.6**	0.26	-4	0.35	94	12.5	0.45	67	0.31	89	32.5	0.06	-78	0.31	90	5.5
	Adj.WGEE	0.00	0.39	95	5.1	-0.11	0.34	93	**6.9**	0.32	0	0.38	94	15.2	0.41	28	0.33	93	25.3	0.22	-31	0.33	93	11.4

* Percentage of missing data for comparator and treatment, respectively; Cov: coverage of 95% CI; Rej: rejection rate; Rel Bias: relative Bias; Est: - log odds ratio; Boldface font indicates that the type I error rate is beyond 2 SEs of simulations.

Table 4: Simulation results under MNAR

		No treatment effect											Moderate treatment effect										
		(45%, 45%)*				(25%, 45%)*							(45%, 45%)*					(25%, 45%)*					(45%, 25%)*
N		Bias	SE	Cov (%)	Rej. (%)	Bias	SE	Cov (%)	Rej. (%)	Est	Rel Bias (%)	SE	Cov (%)	Rej. (%)	Est	Rel Bias (%)	SE	Cov (%)	Rej. (%)	Est	Rel Bias (%)	SE	Cov (%)	Rej. (%)
50	Unadj.logit	0.02	0.41	94	5.8	0.14	0.41	92	**7.9**	0.27	5	0.41	95	10.7	0.15	-44	0.41	93	7.0	0.41	57	0.41	94	18.2
	Adj.logit	0.01	0.45	95	5.2	0.16	0.45	94	**6.2**	0.34	5	0.46	94	12.0	0.18	-43	0.45	94	6.4	0.50	55	0.46	93	19.4
	Unadj.GEE	0.02	0.53	95	5.5	-0.10	0.49	94	**6.1**	0.25	-3	0.53	95	7.7	0.39	48	0.50	94	11.7	0.15	-43	0.50	93	6.2
	Adj.GEE	0.01	0.60	94	**6.4**	-0.10	0.55	93	**6.7**	0.32	1	0.61	95	8.5	0.45	40	0.56	94	11.8	0.22	-33	0.56	94	7.0
	Unadj.GLMM	0.02	0.53	95	4.9	-0.11	0.49	94	5.9	0.25	-4	0.53	94	6.9	0.39	50	0.50	94	11.3	0.14	-47	0.50	93	5.6
	Adj. GLMM	0.01	0.61	94	**6.3**	-0.11	0.56	93	**6.5**	0.33	2	0.61	95	8.4	0.45	42	0.57	94	11.8	0.21	-33	0.57	94	6.6
	Unadj.MI	-0.01	0.50	97	2.8	-0.09	0.47	97	3.4	0.27	4	0.50	98	4.9	0.36	40	0.47	96	9.0	0.19	-26	0.47	97	4.8
	Adj. MI	-0.02	0.57	97	2.6	-0.11	0.53	97	3.3	0.33	3	0.57	98	5.2	0.44	39	0.54	97	9.9	0.24	-26	0.54	97	4.6
	Unadj.WGEE	0.01	0.56	95	4.6	-0.14	0.52	95	5.4	0.25	-3	0.56	96	6.9	0.41	59	0.52	94	11.5	0.11	-57	0.52	94	5.4
	Adj.WGEE	0.01	0.62	95	5.4	-0.11	0.57	94	**6.3**	0.32	2	0.62	95	7.6	0.45	40	0.57	94	11.8	0.21	-34	0.57	95	6.3
200	Unadj.logit	0.00	0.20	95	4.7	0.13	0.20	89	**10.6**	0.29	10	0.20	95	27.3	0.14	-47	0.20	90	11.0	0.41	57	0.20	89	51.1
	Adj.logit	0.00	0.22	94	5.6	0.16	0.22	88	**11.6**	0.34	7	0.22	95	32.6	0.17	-48	0.22	89	12.0	0.49	53	0.22	89	59.1
	Unadj.GEE	0.00	0.26	94	5.9	-0.10	0.24	93	**7.5**	0.26	0	0.26	95	17.2	0.37	41	0.24	93	32.1	0.14	-44	0.25	92	8.8
	Adj.GEE	0.00	0.29	94	5.8	-0.10	0.27	93	**6.6**	0.32	0	0.30	95	19.4	0.42	31	0.27	94	32.2	0.21	-34	0.27	92	11.5
	Unadj.GLMM	0.00	0.26	94	6.0	-0.11	0.24	92	**7.5**	0.26	0	0.26	95	17.2	0.37	41	0.24	93	32.1	0.14	-46	0.25	92	8.2
	Adj. GLMM	0.00	0.29	94	5.8	-0.10	0.27	93	**6.6**	0.32	0	0.30	95	19.3	0.42	31	0.27	94	32.0	0.21	-35	0.28	93	11.2
	Unadj.MI	0.00	0.25	97	2.8	-0.06	0.23	96	4.1	0.27	3	0.25	97	16.2	0.32	23	0.23	96	25.4	0.19	-26	0.23	96	10.1
	Adj. MI	-0.01	0.28	97	2.7	-0.07	0.26	95	4.5	0.32	1	0.28	98	17.9	0.39	21	0.26	96	28.8	0.23	-27	0.27	96	11.0
	Unadj.WGEE	0.00	0.27	95	5.0	-0.13	0.25	92	**7.9**	0.26	0	0.27	96	15.7	0.39	50	0.25	92	33.4	0.12	-53	0.25	91	6.6
	Adj.WGEE	0.00	0.30	95	5.1	-0.10	0.28	94	**6.4**	0.32	0	0.30	96	18.9	0.42	31	0.28	94	31.0	0.21	-33	0.28	94	10.8

* Percentage of missing data for comparator and treatment, respectively; Cov: coverage of 95% CI; Rej: rejection rate; Rel Bias: relative Bias; Est: - log odds ratio; Boldface font indicates that the type I error rate is beyond 2 SEs of simulations.

5. Clinical Trial Examples

The methods tested in the simulation study were also applied to data from five diabetes clinical trials in which an active drug (treatment) was compared with different comparators over 52 weeks. The missing data proportions for the treatment and comparators were 12.5%, 10.2%; 6.6%, 9.3%; 18.0%, 19.5%; 15.6%, 12.8%; and 7.3%, 6.4% for study 1 to 5, respectively. The analyses were conducted on proportion of patients whose 52-week endpoint HbA1c was <7.0%. The LOCF analyses were implemented using logistic models with a factor for treatment only (Unadj.Logit) and with baseline HbA1c (Adj.Logit). The models for the GEE, WGEE, GLMM, and MI analyses included treatment, visit, and treatment by visit interaction, with or without baseline HbA1c. An unstructured covariance matrix or working correlation matrix was used except for WGEE where an independent working correlation matrix was used which was considered as a best fit. Results are displayed in Table 5 and Figure 1. As in the simulated data, covariate adjusted methods yielded larger treatment effects compared to their unadjusted counterparts; and, standard errors were larger and p values smaller from adjusted analyses for all all studies except for study 4, where adjusted and unadjusted methods yielded comparable effects and standard errors. The possible reason is that the baseline HbA1c was not well balanced in study 4 with the treated group associated with a slightly lower value, while the baseline HbA1c was quite well balanced between groups in other studies.

Comment [CHM1]: Was the really the best fit? Seems unlikely.

Table 5. Analysis results from five diabetes clinical studies

Study	Estimates	Unadj. Logit	Adj. Logit	Unadj. GEE	Adj. GEE	Unadj. GLMM	Adj. GLMM	Unadj. MI	Adj. MI	Unadj. WGEE	Adj. WGEE
1	Log odds ratio	0.90	1.20	0.84	1.21	0.84	1.26	0.78	1.06	0.80	1.01
	SE	0.18	0.21	0.19	0.24	0.19	0.25	0.19	0.23	0.20	0.24
	LL	0.54	0.77	0.46	0.73	0.46	0.77	0.41	0.60	0.41	0.54
	UL	1.27	1.62	1.22	1.69	1.22	1.76	1.16	1.53	1.19	1.48
	P	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
2	Log odds ratio	0.95	1.23	0.93	1.22	0.94	1.23	0.93	1.22	0.91	1.22
	SE	0.18	0.21	0.18	0.22	0.19	0.22	0.18	0.21	0.19	0.22
	LL	0.59	0.82	0.56	0.79	0.57	0.80	0.56	0.79	0.53	0.78
	UL	1.31	1.64	1.30	1.65	1.31	1.66	1.29	1.64	1.30	1.65
	P	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
3	Log odds ratio	0.47	0.64	0.43	0.70	0.42	0.98	0.44	0.67	0.45	0.78
	SE	0.18	0.20	0.19	0.27	0.19	0.38	0.18	0.26	0.20	0.27
	LL	0.12	0.24	0.06	0.16	0.04	0.22	0.08	0.15	0.05	0.24
	UL	0.82	1.05	0.81	1.24	0.80	1.75	0.81	1.19	0.85	1.31
	P	0.007	0.002	0.021	0.010	0.026	0.011	0.016	0.010	0.026	0.004
4	Log odds ratio	0.38	0.36	0.45	0.50	0.49	0.50	0.44	0.42	0.50	0.50
	SE	0.17	0.18	0.18	0.19	0.18	0.19	0.17	0.19	0.19	0.20
	LL	0.03	0.00	0.10	0.11	0.13	0.11	0.09	0.05	0.13	0.09
	UL	0.72	0.72	0.81	0.88	0.86	0.88	0.79	0.79	0.88	0.90
	P	0.028	0.047	0.011	0.010	0.007	0.011	0.012	0.025	0.008	0.014
5	Log odds ratio	1.01	1.36	1.09	1.65	1.09	1.74	1.03	1.55	1.07	1.59
	SE	0.17	0.20	0.18	0.23	0.18	0.25	0.17	0.23	0.19	0.23
	LL	0.68	0.97	0.73	1.18	0.73	1.24	0.69	1.09	0.69	1.13
	UL	1.35	1.75	1.45	2.12	1.45	2.24	1.38	2.00	1.45	2.06
	P	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

SE: standard error; LL: lower limit of 95% confidence interval; UL: upper limit of 95% confidence interval;

P: p-value

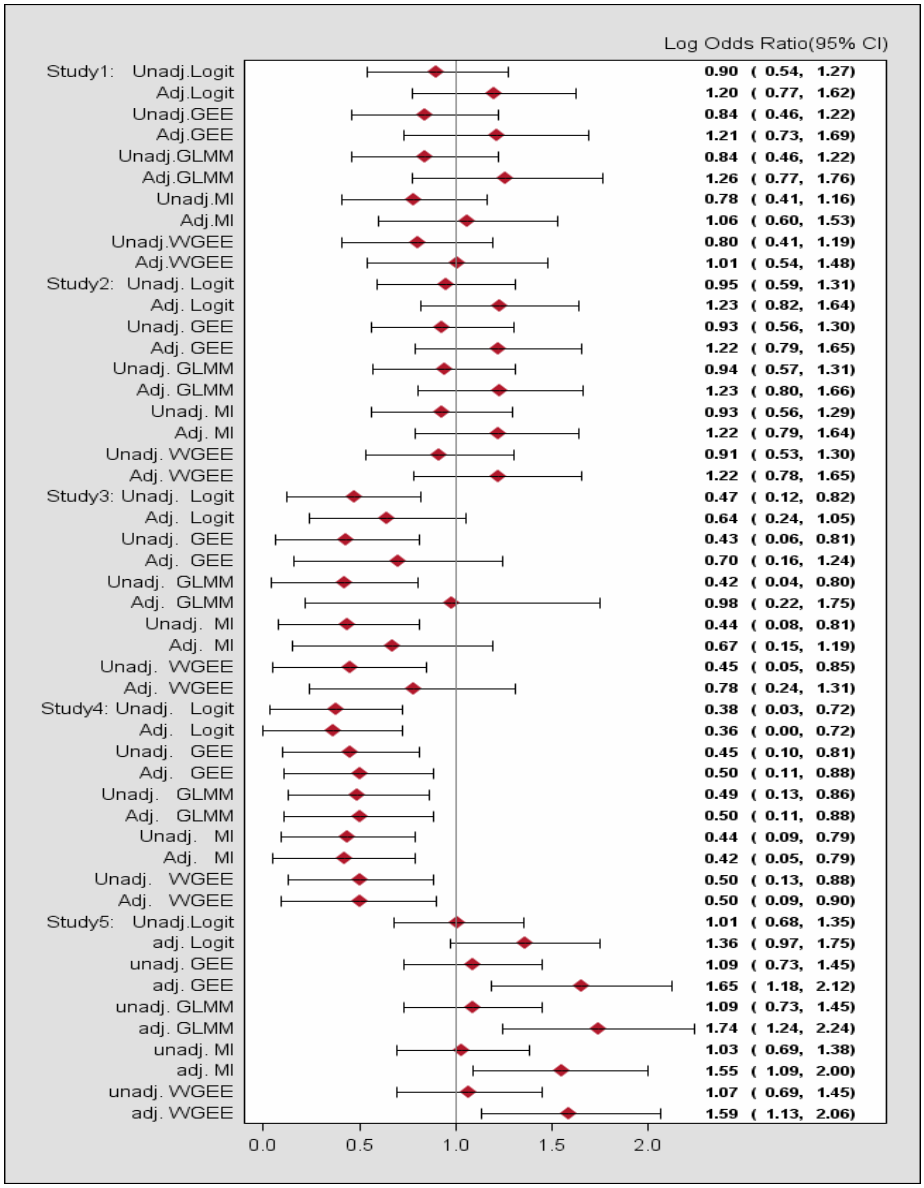

Figure 1. Log odds ratio and 95% CI by study and method

6. Summary and Discussion

In longitudinal clinical trials it is often reasonable to assume missing data arise from an MAR mechanism (Verbeke and Molenberghs 2000; NRC, 2010). Dichotomized continuous outcomes or other binary measures are often of interest. Therefore, appropriate modeling of incomplete longitudinal binary clinical trial data is important.

Simulations were conducted over a variety of scenarios to examine how including or not including baseline severity as a covariate influenced results across several methods for analyzing incomplete longitudinal binary responses. The covariate adjusted analyses generally yielded larger treatment effect estimates and larger standard errors compared with their unadjusted counterpart. The net result of these factors was increased power from the adjusted analyses without increasing type I. Results from five phase 3 diabetes trials were consistent with the simulation findings.

In regards to how the various methods handled missing data, with MNAR data all methods yielded biased results in at least some scenarios and all methods except MI inflated type I error rates in some scenarios. With MCAR data LOCF yielded biased results, inflated type I error rates and had poor CI coverage, whereas results from other methods were not biased. With MAR data, LOCF again yielded biased results. MI and GLMM yielded unbiased results, as expected, because these methods assume MAR. Results from GEE were biased, as expected, because it assumes MCAR. Counter to expectation, results from wGEE were biased in the MAR scenarios where dropout rates differed but were not biased when dropout rates were equal.

Recall the description in section 3.1 of the dropout mechanisms applied to delete data.

In MAR the probability a value (y_i) being missing depended on the observed value at the previous visit (y_{i-1}), expressed as logit (p (y_i missing | y_{i-1})= $a + b * y_{i-1}$. The value b =1.5 was used in the dropout model and the value a was chosen for each treatment group to achieve the desired rates of missing data.

Thus, when the dropout rate was equal in the two treatments they shared a common intercept (a), but when dropout rates differed the intercepts differed. That is, there were separate dropout models for each treatment. However, in the wGEE analyses a single model was used to estimate the weightings for both treatment groups. Therefore, in scenarios where the dropout rates differed, the model used to estimate the weightings was not the same as the model used to generate the missing data. In addition, for MAR logit (p (y_i missing | y_{i-1})= $a + b * y_{i-1}$.

However, the weightings were estimated using logit (p (y_i missing | y_{i-1})= $a + b * y_i$.

The issue of have the same or separate models by treatment group apply to the imputation model in MI. In the present study MI was implemented with separate imputation models for each treatment, thereby accommodating different dropout models for each treatment in the scenarios where dropout rates differed.

These results illustrate the potential importance of modeling considerations in the handling of missing data. For methods that explicitly impute missing values or model dropout, it may be useful to consider separate models for each treatment or for groups of treatments (e.g., all doses of a drug in one group, placebo in the other), especially when rates, timing, and/or reasons for dropout differ.

Although these modeling considerations are important, they should not be taken as motivation to return to use of ad hoc methods. For example, in the MAR scenarios where wGEE yielded type I error rates of 7% – 10%, the corresponding rates from LOCF were 32% - 37%.

The present investigation focused on only one of these many modeling considerations, fitting a single covariate that describes baseline severity. Results support including baseline severity as a covariate in analyses of incomplete longitudinal binary outcomes.

References:

1. ADA. Standards of medical care in diabetes 2013. Diabetes Care 2013;36(Suppl 1):s11–66.
2. Breslow NE, Clayton DG. Approximate inference generalized linear mixed model. Journal of the American Statistical Association 1993; 88:9–25.
3. Daniel Y. T. Fong , Shesh N. Rai & Karen S. L. Lam (2013) Estimating the Effect of Multiple Imputation on Incomplete Longitudinal Data with Application to a Randomized Clinical Study, Journal of Biopharmaceutical Statistics, 23:5, 1004-1022,
4. Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics* 7:1-26.
5. Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
6. Fitzmaurice, G. M., Molenberghs, G., Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative dropouts. J. R. Stat. Soc. 57:691–704.
7. Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73, 13-22.
8. Li, X. M., Mehrotra, D. V., Barnard, J. (2006). Analysis of incomplete longitudinal binary data using multiple imputation. Statistics in Medicine 25:2107–2124.
9. Lipkovich I. Duan Y. Ahmed S. Multiple imputation compared with restricted pseudo-likelihood and generalized estimating equations for analysis of binary repeated measures in clinical studies. Pharm Stat. 2005;4:267-285
10. Little, R. J. A., Rubin, D. B. (1987). Statistical Analysis With Missing Data. New York: Wiley.

11. Liu GF, Zhan XJ. Comparisons of methods for analysis of repeated binary responses with missing data. Journal of Biopharmaceutical Statistics 2011; 21:371–392.
12. Molenberghs G, Kenward MG. (2007), *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.
13. Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M., Mallinckrodt, C., Carrol R. (2004). Analyzing incomplete longitudinal clinical trial data. Biostatistics 5(3): 445–464.
14. National Research Council (2010). The prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
15. Nielsen, S. F. (2003), "Proper and Improper Multiple Imputation," International Statistical Review, 71, 593–607.
16. Preisser JS, Lohman KK, Rathouz PJ. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. Statistics in Medicine 2002; 21:3035 – 3054.
17. Robins, J. M., Rotznitzky, A., Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.* 90:106–121.
18. Rubin, D. B. (1976), "Inference and Missing Data," Biometrika, 63, 581–592.
19. Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

20. Sacks DB, Bruns DE, Goldstein DE, Maclaren NK, McDonald JM, Parrott M. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. Clin Chem 2002;48:436–472
21. SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.
22. Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London, UK: Chapman Hall.
23. Schafer, J. L. (1999). Multiple imputation: A primer. Statistical Methods in Medical Research 8(1):3–15.
24. Shieh, Y.-Y. (2003). Imputation Methods on General Linear Mixed Models of Longitudinal Studies, Federal Committee on Statistical Methodology Research Conference, Arlington, VA.
25. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. Springer: New York, 2000.
26. Wolfinger, R., O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. Journal of Statistical Computation and Simulation 48:233–243.
27. Zeger, S. L. and Liang, K.-Y. (1 986). Longitudinal data analysis for discrete and continuous outcomes. Biometries 42, 121-130.