Published in final edited form as:

Biometrics. 2010 December ; 66(4): 1202–1208. doi:10.1111/j.1541-0420.2010.01405.x.

Multiple Imputation Approaches for the Analysis of Dichotomized Responses in Longitudinal Studies with Missing Data

Kaifeng Lu^{1,*}, Liqiu Jiang², and Anastasios A. Tsiatis³

¹Clinical Biostatistics, Merck Research Laboratories, Rahway, NJ 07065, U.S.A

²Kendle International Inc., Durham, NC 27703, U.S.A

³Department of Statistics, North Carolina State University, Raleigh, NC 27695, U.S.A

Summary

Often a binary variable is generated by dichotomizing an underlying continuous variable measured at a specific time point according to a prespecified threshold value. In the event that the underlying continuous measurements are from a longitudinal study, one can use repeated measures model to impute missing data on responder status as a result of subject drop-out and apply logistic regression model on the observed or otherwise imputed responder status. Standard Bayesian multiple imputation techniques (Rubin, 1987, *Multiple Imputation for Nonresponse in Surveys*) which draw the parameters for the imputation model from the posterior distribution and construct the variance of parameter estimates for the analysis model as a combination of within- and between-imputation variances are found to be conservative. The frequentist multiple imputation approach which fixes the parameters for the imputation model at the maximum likelihood estimates and construct the variance of parameter estimates for the analysis model using the results of (Robins and Wang, 2000, *Biometrika* **87**, 113–124) is shown to be more efficient. We propose to apply (Kenward and Roger, 1997, *Biometrics* **53**, 983–997) degrees-of-freedom to account for the uncertainty associated with variance-covariance parameter estimates for the repeated measures model.

Keywords

Logistic regression; Missing data; Multiple imputation; Repeated measures

1. Introduction

Often a binary response variable is generated by dichotomizing an underlying continuous variable in clinical trials. The dichotomized response is usually a well-established criterion for ascertaining the treatment success against the disease. It can be as important as the underlying continuous measurement from a medical or regulatory point of view. For example, in obesity studies a subject is classified as a 5% weight loss responder if he/she has lost at least 5% of their baseline body weight at the primary time point of interest usually corresponding to the end of the study. According to the FDA 2007 draft guidance document

*kaifeng_lu@merck.com.

Supplementary Materials

Web Appendices, referenced in Section 4, are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

for weight management products, the efficacy of a weight-management product should be assessed by analyses of both mean and categorical changes in body weight. The mean weight loss and the proportion of 5% responders after 1 year of treatment are both considered primary efficacy endpoints and required to be tested at the same significance level. Other examples of responder endpoints include a hemoglobin A1c value below 7% in diabetes trials, at least 50% improvement from baseline in the HAMD-17 total score in depression trials, and an HIV-1 RNA viral load of less than 50 copies per mL in HIV trials.

Therefore, one of the goals of these studies is to model the probability of response as a function of treatment and other covariates at the primary time point of interest usually corresponding to the end of study. A logistic regression model is commonly used for this purpose. Had the measurement been collected at the primary time point from all subjects in the study, the responder status could be ascertained on every subject in the study and the analysis would be routine. It is common, however, in longitudinal clinical trials that some subjects may drop out of the study before reaching the final time point rendering the responder status at the primary time point missing.

A unique feature of dichotomized responses is that the underlying continuous responses are also collected and available for analysis. Since the missing data are more likely to satisfy the missing-at-random (MAR) assumption on the underlying continuous responses than on the dichotomized responses, missing data approaches that rely entirely on the dichotomized responses can be biased. Even when the MAR assumption does hold on the dichotomized responses, missing data methods based on the continuous responses are likely to be more efficient than those based on the dichotomized responses as the continuous responses contain more information than the dichotomized responses. For these reasons, we will only consider missing data methods that utilize the underlying continuous responses in this article.

Multiple imputation (MI) is a convenient and flexible approach for handling missing data and is particularly suitable for analyzing dichotomized responses. The classic MI method (Rubin, 1987, 1996) is based on a Bayesian paradigm where the model parameters are independently drawn from the posterior distribution for each imputed data set, while the frequentist MI method (Wang and Robins, 1998; Robins and Wang, 2000) fixes the model parameters at the maximum likelihood estimates for all imputed data sets. Once the missing data on the underlying continuous responses have been imputed, the observed or otherwise imputed continuous responses are used to construct the dichotomized responses. Each of the completed data sets is then analyzed by standard methods, and the results are later combined to produce estimates and confidence intervals that incorporate missing data uncertainty. The way in which the results are combined across multiple imputed data sets also differs between the Bayesian and frequentist MI methods.

This article will compare various MI methods for the analysis of dichotomized responses and make practical modifications and recommendations.

2. Notation and Models

In a longitudinal study, measurements are collected at multiple, say T , time points. Let the underlying continuous measurement which is used to construct the dichotomized response be denoted by Y_{it} for subject i at time point t for $t = 1, \dots, T$. Without loss of generality, assume that the lower the measurement the better the response. Subject i is classified as a responder at time point t if $Y_{it} < c$ for some cutoff value c . Let $\{Z_{i1}, \dots, Z_{iT}\}$ denote the induced responder statuses over time for subject i , i.e., $Z_{it} = I\{Y_{it} < c\}$. Consider a randomized clinical trial with two treatment groups, A and B , where A denotes the active treatment and B the control. Let G_i denote the treatment (either A or B) for subject i .

The main objective is to evaluate the treatment effect on the responder status at the primary time point of interest, $t = T$. Ordinarily, a logistic regression model is used to estimate the effects of covariates on a binary response. Specifically, the logistic regression model assumes

$$\text{logitPr}(Z_{iT}=1)=\mathbf{W}_i'\boldsymbol{\beta}, \quad (1)$$

where the ‘logit’ of a probability p is defined as $\log\{p/(1-p)\}$, $\mathbf{W}_i$ is a column vector constructed from baseline covariates including the treatment indicator, and $\boldsymbol{\beta}$ the associated regression coefficients. In particular, if the treatment is the only effect in the model, $\mathbf{W}_i = (1, I(G_i = A))'$. If adjustment for baseline measurement of the underlying continuous variable is to be made, $\mathbf{W}_i = (1, I(G_i = A), Y_{i1})'$. In either case, the parameter of interest is the log odds ratio corresponding to the treatment indicator $I(G_i = A)$.

It is common in longitudinal clinical trials that some subjects may drop out of the study before reaching the final time point rendering the responder status at the primary time point missing. According to our strategy, a fitted repeated measures model would be used as an imputation model for the missing data. Such a repeated measures model for the continuous response vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT})'$ assumes

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\psi}, \mathbf{V}_i), \quad (2)$$

where $\mathbf{X}_i$ is the design matrix for subject i based on baseline covariates and time, $\boldsymbol{\psi}$ the corresponding regression coefficients, $\mathbf{V}_i = \mathbf{V}_i(\boldsymbol{\sigma})$ the covariance matrix for $\mathbf{Y}_i$, and $\boldsymbol{\sigma}$ the variance-covariance parameters for the repeated measures model. In the presence of missing data, the full response vector $\mathbf{Y}_i$ can be partitioned into the observed components $\mathbf{Y}_i^{\text{obs}}$ and the unobserved components $\mathbf{Y}_i^{\text{mis}}$. Let $\mathbf{X}_i^{\text{obs}}$ and $\mathbf{V}_i^{\text{obs}}$ denote the design matrix and covariance matrix corresponding to $\mathbf{Y}_i^{\text{obs}}$; $\mathbf{X}_i^{\text{mis}}$ and $\mathbf{V}_i^{\text{mis}}$ the design matrix and covariance matrix corresponding to $\mathbf{Y}_i^{\text{mis}}$; $\mathbf{V}_i^{\text{obs,mis}}$ the covariance matrix between $\mathbf{Y}_i^{\text{obs}}$ and $\mathbf{Y}_i^{\text{mis}}$, and $\mathbf{V}_i^{\text{mis,obs}} = \mathbf{V}_i^{\text{obs,mis}}'$. The repeated measures model (2) implies that $\mathbf{Y}_i^{\text{obs}} \sim N(\mathbf{X}_i^{\text{obs}}\boldsymbol{\psi}, \mathbf{V}_i^{\text{obs}})$, and $\mathbf{Y}_i^{\text{mis}} | \mathbf{Y}_i^{\text{obs}} \sim N(\mathbf{X}_i^{\text{mis}}\boldsymbol{\psi} + \mathbf{V}_i^{\text{mis,obs}}\mathbf{V}_i^{\text{obs-1}}(\mathbf{Y}_i^{\text{obs}} - \mathbf{X}_i^{\text{obs}}\boldsymbol{\psi}), \mathbf{V}_i^{\text{mis|obs}})$, independent across $i = 1, \dots, n$, where $\mathbf{V}_i^{\text{mis|obs}} = \mathbf{V}_i^{\text{mis}} - \mathbf{V}_i^{\text{mis,obs}}\mathbf{V}_i^{\text{obs-1}}\mathbf{V}_i^{\text{obs,mis}}$.

In particular, if the constrained longitudinal data analysis (cLDA) model (Liang and Zeger, 2000; Fitzmaurice, Laird, and Ware, 2004, strategy 2 on page 126; Liu, et al, 2009) is used,

we have $\mathbf{X}_i = \begin{pmatrix} \mathbf{I}_T & \mathbf{0}_{1 \times (T-1)} \\ & \mathbf{I}_{T-1} \end{pmatrix}$ for $G_i = A$, and $\mathbf{X}_i = (\mathbf{I}_T, \mathbf{0}_{T \times (T-1)})$ for $G_i = B$. Correspondingly,

$\boldsymbol{\psi} = (\boldsymbol{\mu}_B', \boldsymbol{\delta}')'$, where $\boldsymbol{\mu}_B$ is a vector of length T representing the means for group B , and $\boldsymbol{\delta}$ a vector of length $T - 1$ representing the mean differences at postbaseline time points between groups A and B . The baseline mean is constrained to be the same across treatment groups due to randomization. The above formulation of the cLDA model includes treatment as the only effect and has a saturated mean profile and unstructured covariance. This is reminiscent of the MMRM (mixed-effects model repeated measures) model proposed by Mallinckrodt, Clark and David (2001), which is particularly suitable for analyzing clinical trial data since subjects in clinical trials are often evaluated at a fixed and relative small number of time points (Siddiqui, Hung and O'Neill, 2009).

If the treatment is the only effect in the logistic regression model, (1) and (2) imply that $\Pr(Z_{iT} = 1|G_i) = \{1 + \exp(-\beta_1 - \beta_2 I(G_i = A))\}^{-1} = \Phi((c - \psi_T - \psi_{2T-1} I(G_i = A))/\sigma_T)$, where β_j denotes the j th component of β , Φ the cumulative distribution function of the standard normal distribution, ψ_j the j th component of ψ , and σ_T the standard deviation of Y_{iT} . Consequently, β is defined by the parameters of the distribution of Y_{iT} .

On the other hand, if the baseline measurement of the underlying continuous variable is adjusted in the logistic regression model, (1) implies $\Pr(Z_{iT} = 1|G_i, Y_{i1}) = \{1 + \exp(-\beta_1 -$

$$\beta_2 I(G_i = A) - \beta_3 Y_{i1})\}^{-1}, \text{ while (2) implies } \Pr(Z_{iT} = 1|G_i, Y_{i1}) = \Phi\left(\frac{c - (\psi_T + \psi_{2T-1} I(G_i = A) + (\rho\sigma_T / \sigma_1)(Y_{i1} - \psi_1))}{\sigma_T(1 - \rho^2)^{1/2}}\right),$$

where σ_1 denotes the standard deviation of Y_{i1} , and ρ the correlation between Y_{i1} and Y_{iT} . Consequently, the two models are not compatible as they imply different functional forms for the response probabilities as a function of random variable Y_{i1} . Therefore, the first question we need to address is what a logistic regression model and subsequent analysis of the responder data is actually estimating for this incompatibility. The point of view taken in this paper is that the logistic regression model is routinely used in practice and understanding the theoretical properties of the resulting estimator is important even if the model is not specified correctly. If the true relationship for the probability of response as a function of covariates does not deviate greatly from that of a logistic regression model, then the parameter estimates from a logistic regression model will converge to a model that most closely approximates the true relationship (i.e., the least false parameter). Such inference will still be useful if the approximation is reasonable. In practical terms, both on goodness of fit grounds and on the grounds of individual parameter estimate comparisons (logit coefficients are about 1.8 times the size of otherwise comparable probit results), there is very little difference between the logistic and probit regression procedures (provided that there are not too many observations at extreme probability values) (Aldrich and Nelson, 1984, pp. 64–65).

3. Imputation Approaches

3.1 Bayesian imputation procedure

The Bayesian imputation procedure imputes the missing data from the posterior predictive distribution of the missing data under some complete-data model and prior, $p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}) = \int p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \theta)p(\theta|\mathbf{Y}_{\text{obs}})d\theta$, where $\theta = (\psi, \sigma)$ for the repeated measures model (2). This suggests that we first draw θ from the posterior distribution, $p(\theta|\mathbf{Y}_{\text{obs}})$, and then draw $\mathbf{Y}_{\text{mis}}$ from the conditional distribution, $p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \theta)$, evaluated at the value of θ obtained in the first step. This strategy works if the posterior distribution, $p(\theta|\mathbf{Y}_{\text{obs}})$, is easy to simulate from, e.g., by using the regression method (Rubin, 1987) when the missing data exhibit a monotone pattern. Otherwise, the missing data may be imputed using the Markov Chain Monte Carlo (MCMC) algorithm (Schafer, 1997; Little and Rubin, 2002). Both algorithms have been implemented in SAS PROC MI.

Suppose the missing data are filled in m times to generate m complete data sets. The observed or otherwise imputed continuous responses are used to construct the dichotomized responses. Each of the m completed data sets is analyzed by the logistic regression model at time point T , yielding m estimates of the log odds ratio and associated variance. The results from the m complete data sets are combined using standard MI techniques (Rubin, 1987). This step has been implemented in SAS PROC MIANALYZE.

Despite its conceptual simplicity and flexibility, the above MI procedure is not valid for the analysis of dichotomized responses because Rubin's variance estimator is biased when the analysis model is different from the imputation model (Meng, 1994; Robins and Wang,

2000). This is true even when the imputation and analysis models are compatible, e.g. when the treatment is the only effect in the logistic regression model.

3.2 Frequentist imputation procedure

Wang and Robins (1998) study the statistical properties of MI estimators from a frequentist point of view using large sample theory. They show that the imputation procedure which fixes the parameters at the maximum likelihood estimates is more efficient than the one which draws the parameters from the posterior distribution. Robins and Wang (2000) derive the large sample properties of MI estimators when the imputation and analysis models may be different or incompatible. For the analysis of dichotomized responses, the repeated measures model (2) serves as the imputation model, while the logistic regression model (1) serves as the analysis model.

Let $\hat{\psi}$ and $\hat{\sigma}$ denote the maximum likelihood (ML) and restricted maximum likelihood (REML) estimates of ψ and σ for the repeated measures model. The frequentist imputation method is based on the conditional distribution of the missing data given the observed data evaluated at the ML estimate of the parameters in the imputation model, $Y_i^{\text{mis}} | Y_i^{\text{obs}}, \hat{\psi}$, $\hat{\sigma} \sim N(X_i^{\text{mis}} \hat{\psi} + V_i^{\text{mis, obs}} V_i^{\text{obs-1}} (Y_i^{\text{obs}} - X_i^{\text{obs}} \hat{\psi}), V_i^{\text{mis|obs}})$, where $V_i = V_i(\hat{\sigma})$. Let Y_{ij}^{mis} denote the j th imputation of Y_i^{mis} , and $Y_{ij} = (Y_i^{\text{obs}}, Y_{ij}^{\text{mis}})$, $j = 1, \dots, m$.

Once the missing data on the underlying continuous responses have been filled in, the observed or otherwise imputed continuous responses can be used to construct the dichotomized responses. Let $Z_i = Z(Y_i) = I(Y_i < c)$ denote the dichotomized responses based on the full data on subject i , and $Z_{ij} = Z(Y_{ij})$ the corresponding dichotomized responses in the j th imputed data set. Let $\hat{\beta}_j$ denote the ML estimate of β from the logistic regression model on the j th imputed data set. The combined point estimate for β from MI is the average of m complete-data estimates, $\hat{\beta} = m^{-1} \sum_j \hat{\beta}_j$.

Robins and Wang (2000) give the asymptotic variance of $n^{1/2}(\hat{\beta} - \beta)$ in Theorem 1 and a consistent estimator of the asymptotic variance in Corollary 1. Literal application of their results would entail the evaluation of the score function with respect to the variance-covariance parameters, σ , for the repeated measures model, which has a highly complex analytic form and is usually not output by off-the-shelf software packages. We propose to exclude σ from the set of parameters for the repeated measures model by fixing it at $\hat{\sigma}$, i.e. ψ will be the only parameter for the repeated measures model in the application of Robins-Wang variance estimator.

The proposed modification is similar to the common practice in linear models, where the variance-covariance parameters are assumed known when deriving the variance for the mean parameter estimates, $\text{var}(\hat{\psi}) = (X'V(\sigma)^{-1}X)^{-1}$. An estimate of this variance is then obtained by substituting $\hat{\sigma}$ for σ , $\widehat{\text{var}}(\hat{\psi}) = (X'V(\hat{\sigma})^{-1}X)^{-1}$. This may underestimate the true variability of $\hat{\psi}$ because the uncertainty associated with the estimation of σ is not taken into account. Kenward and Roger (1997) provide an approximation that involves inflating the above naïve estimate of $\text{var}(\hat{\psi})$ and computing Satterthwaite-type degrees of freedom. We propose to use the Kenward-Roger degrees of freedom associated with $\hat{\psi}$ to make inference on β , i.e., the inference on β will be based on the t distribution with Kenward-Roger degrees of freedom instead of the asymptotic normal distribution. In particular, the degrees of freedom for the estimated log odds ratio between groups A and B at time point T will be the same as that for the estimated mean difference between groups A and B at time point T .

3.3 Hybrid imputation procedure

By fixing the parameters at the ML estimates, the frequentist imputation procedure gains efficiency over the Bayesian imputation procedure, but at the cost of increased complexity in calculating the estimated variance for the resulting MI estimator. A hybrid imputation procedure imputes the missing data using the frequentist method but combines the analysis results using the Bayesian method. To further simplify the imputation of missing data on the underlying continuous responses, we can impute each missing value using a univariate normal distribution with mean equal to the predicted value at the corresponding time point and variance equal to the squared standard error for the prediction. This is straightforward to implement as the predicted values and associated standard errors are readily available from off-the-shelf software packages such as SAS PROC MIXED.

Once the missing data on the underlying continuous responses have been filled in, the observed or otherwise imputed continuous responses can be used to construct the dichotomized responses. Each of the m completed data sets is analyzed by the logistic regression model, yielding m estimates of the log odds ratio and associated variance. The results from the m complete data sets are combined using Bayesian MI procedure.

4. Simulation Studies

4.1 Data generation

Simulation studies are conducted to evaluate the performance of the 3 imputation procedures. Data for the underlying continuous responses were simulated from multivariate normal distributions for two treatment groups with equal allocation and T repeated time points per subject. Let μ_j denote the mean vector for repeated measurements on a subject in group $j = A, B$, and Σ the covariance matrix. We considered $T = 4$ for moderate number of time points and $T = 7$ for large number of time points. The mean vectors and covariance matrices for the two groups under the alternative hypothesis are $\mu_A = (3.0, 2.0, 1.5, 0.7)'$, $\mu_B = (3.0, 2.5, 2.3, 2.0)'$, and

$$\Sigma = 4 \begin{pmatrix} 1 & 0.7 & 0.4 & 0.2 \\ 0.7 & 1 & 0.7 & 0.4 \\ 0.4 & 0.7 & 1 & 0.7 \\ 0.2 & 0.4 & 0.7 & 1 \end{pmatrix}$$

for $T = 4$; $\mu_A = (3.0, 2.5, 2.0, 1.7, 1.5, 1.1, 0.7)'$, $\mu_B = (3.0, 2.7, 2.5, 2.4, 2.3, 2.1, 2.0)'$, and

$$\Sigma = 4 \begin{pmatrix} 1 & 0.84 & 0.70 & 0.53 & 0.40 & 0.28 & 0.20 \\ 0.84 & 1 & 0.84 & 0.70 & 0.53 & 0.40 & 0.28 \\ 0.70 & 0.84 & 1 & 0.84 & 0.70 & 0.53 & 0.40 \\ 0.53 & 0.70 & 0.84 & 1 & 0.84 & 0.70 & 0.53 \\ 0.40 & 0.53 & 0.70 & 0.84 & 1 & 0.84 & 0.70 \\ 0.28 & 0.40 & 0.53 & 0.70 & 0.84 & 1 & 0.84 \\ 0.20 & 0.28 & 0.40 & 0.53 & 0.70 & 0.84 & 1 \end{pmatrix}$$

for $T = 7$. The covariance matrices are chosen to have decreasing correlations over time with an average correlation of 0.5 for $T = 4$ and 0.6 for $T = 7$ (such values are reasonable for clinical trials, Frison and Pocock (1992)) and with an effect size (δ/σ) of 0.65 at the last time point. The parameter values for $T = 4$ represent a subset sampling of $T = 7$.

The distributions for the two groups are identical under the null hypothesis with the mean vector for *A* equal to that for group *B*. We considered $n = 100$ (i.e., 50 per group) for moderate sample sizes and $n = 200$ (i.e., 100 per group) for large sample sizes.

The dichotomized responses were constructed as $Z_{it} = I(Y_{it} < c)$ for some cutoff value c . We considered $c = 1.8$ for high response rates and $c = 0.5$ for low response rates. Specifically, the response rates at the last time point under the alternative hypothesis were 0.71 for group *A* and 0.46 for group *B* for $c = 1.8$, and 0.46 for group *A* and 0.23 for group *B* for $c = 0.5$. The true log odds ratio in response rate at the last time point between the two groups was 1.05 for $c = 1.8$ and 1.07 for $c = 0.5$. When the baseline of the underlying continuous response is adjusted for in the logistic regression model, the estimated log odds ratios from the MI approaches were compared with results from the logistic regression analysis without missing data.

In general, the monotone missing data occur most frequently in clinical trials as a result of early dropouts. Therefore, for simplicity, we considered the case of monotone missing data. The following missing data models for the underlying continuous response were used in the simulation studies: $\{Y_{it}, \dots, Y_{iT}\}$ were set to missing

- with probability, $\{1 + \exp(4.3 - 0.7Y_{i,t-1})\}^{-1}$, to mimic missing data due to lack of efficacy, or
- with probability, $\{1 + \exp(4.3 - 0.7Y_{i,t-1} - 0.6I(G_i = A))\}^{-1}$, to mimic missing data due to lack of efficacy and adverse event, or
- if $Y_{i,t-1} > 4.5$ to mimic missing data due to initiation of rescue therapy.

The missing data generated through the above models are clearly missing at random. The proportion of missing data at the last time point ranged from 28% to 61% for group *A*, and from 32% to 47% for group *B*. The percentage of missing data for each time point under different scenarios can be found in the web-based supplementary materials.

Table 1 shows the combinations of sample sizes, number of time points, missing data generation, response rates, and baseline adjustment used in the simulation studies.

4.2 Results

The imputation approaches were compared in terms of empirical type I error and power of the test of treatment effect, bias of the point estimate of log odds ratio, and coverage probability of the 95% confidence interval (CI). Results were based on 5,000 repetitions and $m = 10$ imputations for each simulation scenario. With 5,000 repetitions for each simulation scenario, the 95% CI of the empirical type I error rate is (0.044, 0.056), and the 95% CI of the empirical coverage of 95% CI is (0.944, 0.956).

Table 2 presents the empirical type I error under the null hypothesis, power, relative bias of log odds ratio estimate and coverage of 95% CI under the alternative hypothesis. The results indicate that all MI methods controlled the type I error. The Bayesian MI method was the most conservative, the hybrid MI method was less conservative, and the frequentist MI method most properly controlled the type I error and was the most powerful. The power loss of the Bayesian MI method was substantial for small to moderate sample sizes. The bias generally appeared to be negligible for all MI methods. The coverage probabilities for the Bayesian and hybrid MI methods were above the target level, while those for the frequentist MI method seemed to track the target level.

The asymptotically unbiased variance estimator for the analysis of dichotomized data is the Robins-Wang variance estimator. To reduce the complexity of implementation, the

variance-covariance parameters in the repeated measures imputation model have been assumed known and fixed at the (restricted) maximum likelihood estimates. However, the Robins-Wang variance estimator tends to underestimate the true variability in this case, which, in turn, can lead to inflated type I error for small samples. For example, the type I errors from the simulation studies were 5.4, 6.1, 5.6, 5.2, 5.8, 6.1, 5.5, 5.4, 5.3 and 6.4 per cent for cases 1 through 10 (i.e. cases with 50 subjects per group), all above 5% and with about a half exceeding 5.6%, the upper bound of 95% CI for the target type I error rate of 5%. By applying the Kenward-Roger adjustment for the degrees of freedom, the frequentist imputation approach seemed to control the type I error and maintain the confidence interval coverage probability.

5. Example

We applied the MI approaches to the data set from a long-term weight-loss study (ClinicalTrials.gov identifier: NCT00092859) which motivated this research. This was a 52-week, multicenter, randomized, double-blind, placebo-controlled study to examine the safety and efficacy of MK-0557, a potent, highly selective, orally active neuropeptide Y5 receptor (NPY5R) antagonist to induce weight loss in overweight and obese patients. The study included a 2-week diet and exercise single-blind placebo run-in prior to randomization. A total of 1661 patients were randomized in a 1:2 ratio to receive placebo or MK-0557 1 mg once daily for 52 weeks. The primary hypothesis was that MK-0557 was more effective than placebo at inducing weight loss after 52 weeks of treatment. The primary analyses were based on the mean change from baseline (week 0) in body weight at week 52, the proportion of patients who lost 5% of their baseline body weight (5% responders) at week 52, and the proportion of patients who lost 10% of their baseline body weight (10% responders) at week 52. The study results were reported in Erond, et al (2006).

For illustration, we focused on the 5% responders analysis. Only 854 patients had body weight data at week 52 (approximately 54% of patients randomized to placebo and 55% of patients randomized to MK-0557). To impute the missing data at week 52, a longitudinal analysis model was fitted to the repeated measures (at weeks 4, 8, 12, 16, 20, 24, 28, 36, 44 and 52) of percent change from baseline in body weight. The model included terms for baseline weight, weight lost during the 2-week run-in, study center, treatment, week of visit as a categorical variable, and the interactions for week-by-weight loss during run-in and week-by-treatment. An unstructured variance-covariance model was used for the repeated measures. Once the missing data on percent change from baseline in body weight were imputed, each patient was classified as a 5% weight loss responder or nonresponder at week 52, and a logistic regression model was fitted to the observed or otherwise imputed 5% responder statuses in each imputed data set. The logistic regression model included terms for baseline body weight, weight lost during the 2-week run-in, study center, and treatment. The results from 10 imputed data sets were combined to produce estimates and confidence intervals for the odds ratio between MK-0557 and placebo.

Each of the 3 imputation approaches discussed in Section 3 was applied to the analysis of 5% responders. For the Bayesian imputation approach, we used the following Jeffrey's prior, $p(\boldsymbol{\psi}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-11/2}$, where $\boldsymbol{\psi}$ is the vector of regression parameters and $\boldsymbol{\Sigma}$ the variance-covariance matrix in the repeated measures model for percent change from baseline in body weight. Since the missing data were not monotone, we used an MCMC algorithm based on the Gibb's sampler, applied in three steps consisting of an imputation step for $\mathbf{Y}_{\text{mis}}$ and two conditional posterior steps for drawing the values of $\boldsymbol{\psi}$ and $\boldsymbol{\Sigma}$ (Little and Rubin, 2002, Example 10.3, pp. 204–205). A single chain was used with 500 burn-in iterations, 1000 MCMC iterations (excluding the burn-in iterations), and 100 iterations between imputations. The ML estimate $\hat{\boldsymbol{\psi}}$ for $\boldsymbol{\psi}$ and REML estimate $\hat{\boldsymbol{\Sigma}}$ for $\boldsymbol{\Sigma}$ from the repeated measures model

were used as initial values to begin the MCMC process. The convergence of the chain was verified by the trace plots. For the hybrid imputation approach, the predicted values and standard errors from SAS PROC MIXED were used to impute missing values at each time point. For the frequentist imputation approach, missing data were imputed based on the conditional distribution given the observed data evaluated at $\psi = \hat{\psi}$ and $\Sigma = \hat{\Sigma}$. To combine results on log odds ratio from 10 imputed data sets, SAS PROC MIANALYZE (Rubin's variance estimator) was used for the Bayesian and hybrid imputation approaches, while Robins-Wang's variance estimator with Kenward-Roger degrees of freedom was used for the frequentist imputation approach.

Table 3 displays the results on treatment effect in terms of odds ratio (OR) and logarithm of odds ratio (Log OR) for different methods. We observed that treatment effect was statistically significant with each method, indicating that there was some evidence that the odds of 5% responders in the MK-0557 group at one year was higher than that of the placebo group. Consistent with the results from the simulation studies, the frequentist imputation approach produced smallest standard error and shortest confidence interval. The analysis was carried out using SAS 9.1 and 557011 was used as the initial random seed for each imputation method.

With 1661 randomized subjects (1555 with both a baseline measurement and at least one postbaseline measurement), the study was overpowered to detect small treatment differences in weight loss. To better illustrate the differences among the three imputation approaches, we performed the same analysis among the female patients (1343 with both a baseline measurement and at least one postbaseline measurement, 906 in the MK-0557 group and 437 in the placebo group). The results are also reported in Table 3. The frequentist imputation approach produced smallest standard error and statistically significant treatment difference, while the Bayesian imputation approach produced largest standard error (24% larger than the frequentist approach) and non-statistically significant treatment difference, although the magnitude of treatment difference was similar between the two methods.

For the subgroup analysis among the male patients, the logistic regression model did not converge due to the small sample sizes and the inclusion of study center as a categorical variable in the model. After dropping study center from the logistic regression model, Table 3 shows that none of three imputation methods produced statistically significant treatment differences due to the small sample sizes. The standard error of the frequentist imputation approach was again smaller than those of the other two imputation approaches.

Of note, although statistically significant at 52 weeks based on all patients, the magnitude of MK-0557 induced weight loss was not clinically meaningful and the development program was subsequently terminated.

6. Discussion

The Bayesian multiple imputation approach has been shown to be conservative for the analysis of dichotomized data as Rubin's variance estimator tends to overestimate the true variability. The simplified hybrid multiple imputation approach, which uses predicted values and associated standard errors for imputation and Rubin's variance estimator for inference, is also conservative for the analysis of dichotomized data. Therefore, despite their intuitive appeal and computational simplicity, these two methods should be discouraged in this context if power and efficiency are important considerations. The proposed frequentist multiple imputation approach, which fixes the variance-covariance parameters at the (restricted) maximum likelihood estimates and applies the Kenward-Roger adjustment of degrees of freedom for inference, appears to be able to properly control the type I error and

is the most powerful among the three imputation approaches evaluated. We recommend this method for the analysis of dichotomized data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The first author is grateful to Drs. Devan Mehrotra, Guanghan (Frank) Liu, and many other Merck colleagues for their constructive comments. The authors would like to thank the associate editor for the many helpful comments and suggestions that led to significant improvement of the paper. This research was supported by NIH grants; contract/grant number: R37-AI31789 and R01-CA51962.

References

- Aldrich, JH.; Nelson, FD. Linear Probability, Logit, and Probit Models. Newbury Park, CA: SAGE Publications, Inc; 1984.
- Erondur N, Gantz I, Musser B, Suryawanshi S, et al. Neuropeptide Y5 receptor antagonism does not induce clinically meaningful weight loss in overweight and obese adults. *Cell Metabolism*. 2006; 4:275–282. [PubMed: 17011500]
- Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied Longitudinal Analysis. Hoboken, New Jersey: Wiley; 2004.
- Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine*. 1992; 11:1685–1704. [PubMed: 1485053]
- Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*. 1997; 53:983–997. [PubMed: 9333350]
- Liang KY, Zeger S. Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhyā: The Indian Journal of Statistics, Series B*. 2000; 62:134–148.
- Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2. New York: Wiley; 2002.
- Liu G, Lu K, Mogg R, Mallick M, Mehrotra DV. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? *Statistics in Medicine*. 28:2509–2530. [PubMed: 19610129]
- Mallinckrodt CH, Clark WS, David SR. Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics*. 2001; 11:9–21. [PubMed: 11459446]
- Meng XL. Multiple imputation inferences with uncongenial sources of input. *Review of Economics and Statistics*. 1994; 76:591–608.
- Robins JM, Wang N. Inference for imputation estimators. *Biometrika*. 2000; 87:113–124.
- Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
- Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association*. 1996; 91:473–489.
- Schafer, JL. Analysis of Incomplete Multivariate Data. New York: Chapman & Hall; 1997.
- Siddiqui O, Hung JHM, O'Neill R. MMRM vs. LOCF: a comprehensive comparison based on simulation study and 25 NDA datasets. *Journal of Biopharmaceutical Statistics*. 2009; 19:227–246. [PubMed: 19212876]
- Wang N, Robins JM. Large sample theory for parametric multiple imputation procedures. *Biometrika*. 1998; 85:935–948.

Table 1

Simulation scenarios.

Case	#subj./group	#time points	Missing data	Resp. rates	Baseline
1	50	4	lackeff	high	no
2	50	4	lackeff+AE	high	no
3	50	4	rescue	high	no
4	50	4	lackeff	low	no
5	50	4	lackeff	high	yes
6	50	7	lackeff	high	no
7	50	7	lackeff+AE	high	no
8	50	7	rescue	high	no
9	50	7	lackeff	low	no
10	50	7	lackeff	high	yes
11	100	4	lackeff	high	no
12	100	4	lackeff+AE	high	no
13	100	4	rescue	high	no
14	100	4	lackeff	low	no
15	100	4	lackeff	high	yes
16	100	7	lackeff	high	no
17	100	7	lackeff+AE	high	no
18	100	7	rescue	high	no
19	100	7	lackeff	low	no
20	100	7	lackeff	high	yes

Table 2

Simulation results (Bayes: Bayesian ML, Hybrid: hybrid ML, Freq: frequentist ML).

Case	Type I Error (%)			Power (%)			Relative Bias (%)			Coverage (%)		
	Bayes	Hybrid	Freq	Bayes	Hybrid	Freq	Bayes	Hybrid	Freq	Bayes	Hybrid	Freq
1	3.3	3.9	4.9	57.5	60.7	66.3	3.4	4.4	4.9	96.9	96.4	94.6
2	3.4	4.2	5.3	52.4	58.8	65.4	3.6	4.6	5.5	97.0	96.2	94.2
3	3.5	4.3	4.8	56.2	61.1	65.0	2.2	3.7	4.1	97.2	96.7	95.4
4	2.7	3.1	4.7	58.1	61.3	64.6	2.8	3.8	4.2	96.8	96.4	95.3
5	3.4	4.0	5.1	58.1	61.7	67.7	0.6	1.4	1.9	96.8	96.4	94.5
6	2.7	4.1	4.8	45.7	54.5	60.6	-0.0	2.9	3.5	97.4	96.2	94.7
7	2.2	4.4	4.2	36.4	49.9	55.3	-1.7	2.4	3.6	97.9	96.2	94.9
8	3.2	4.2	4.7	51.3	57.3	62.2	-0.1	2.5	2.9	97.4	96.6	95.4
9	2.1	3.1	4.5	48.2	54.7	58.2	0.0	2.8	3.3	97.2	96.6	95.2
10	2.7	4.3	5.2	46.5	55.2	61.8	-0.9	1.8	2.4	97.1	96.5	94.8
11	2.8	3.5	4.6	87.7	89.1	92.5	0.5	1.0	1.2	97.0	96.5	95.1
12	2.8	3.9	4.7	84.9	87.4	91.0	0.9	1.7	1.7	97.2	96.6	94.4
13	3.2	4.0	5.0	86.9	88.8	91.1	0.3	1.2	1.5	96.9	96.3	94.7
14	2.9	3.2	4.4	88.0	89.2	91.5	0.3	0.6	0.8	96.5	96.2	95.2
15	2.9	3.6	4.7	88.4	90.0	92.9	-0.2	0.2	0.4	97.1	96.7	95.0
16	3.3	4.1	4.9	81.1	86.2	89.5	0.3	1.9	2.1	97.2	96.3	94.3
17	2.6	4.6	4.6	74.6	83.0	87.0	0.3	2.0	2.5	97.5	96.1	94.3
18	3.3	4.3	4.8	85.2	88.3	90.9	0.0	1.3	1.5	97.1	96.3	95.0
19	2.6	3.1	4.8	83.6	86.6	88.5	0.7	2.0	2.4	96.7	95.6	94.5
20	3.2	4.3	5.0	81.9	87.0	90.2	-0.6	0.9	-2.4	97.3	96.3	95.1

Results of estimates for treatment effect (OR: odds ratio, Log OR: logarithm of odds ratio) for the weight-loss study. Study center was excluded from the logistic regression model for the subgroup analysis among male patients due to convergence issues.

Table 3

Population	MI method	Estimate (Log OR)	Std Error	t value	d.f.	Pr > t	Estimate (OR), 95% CI
All patients	Bayesian	0.347	0.121	2.866	473.17	0.004	1.42 (1.12, 1.80)
	Hybrid	0.345	0.138	2.502	79.49	0.014	1.41 (1.07, 1.86)
	Frequentist	0.307	0.113	2.715	1298.84	0.007	1.36 (1.09, 1.70)
Female patients	Bayesian	0.299	0.153	1.962	68.84	0.054	1.35 (0.99, 1.83)
	Hybrid	0.305	0.145	2.100	102.29	0.038	1.36 (1.02, 1.81)
	Frequentist	0.290	0.123	2.352	1112.48	0.019	1.34 (1.05, 1.70)
Male patients	Bayesian	0.411	0.322	1.276	374.53	0.203	1.51 (0.80, 2.84)
	Hybrid	0.429	0.331	1.299	228.29	0.195	1.54 (0.80, 2.95)
	Frequentist	0.395	0.291	1.355	176.49	0.177	1.48 (0.84, 2.64)