A Simulation Study to Compare Multiple Imputation Methods under Missing Not at Random Assumption

Jianjun (David) Li[1], Lingfeng Yang[2]
[1]Pfizer, Inc., 500 Arcola Road, Collegeville, PA, 19426
[2]BMS, 311 Pennington Rocky Hill Rd, Pennington, NJ 08534

Abstract

Because missing values are neither observable nor depend on observed values, data missing not at random (MNAR) poses unique challenges in data analysis. A simulation study that compares four multiple imputation (MI) methods under MNAR assumption was conducted to address regulatory concerns of missing data in a clinical trial. The four MI methods for comparison are jumping to control (JC), coping difference from control (CDC), imputation with observed means in each arm (GM), and last z-value carried forward (LZCF). A variety of scenarios of missing data proportions in drug and placebo arms was considered to evaluate these methods in terms of power and type I error rate. The simulation study shows that (1) CDC performs best among the four MI methods; (2) Intuitively conservative method, JC, does not necessarily protect against type I error rate better than CDC; (3) Analysis results are not sensitive to the number of imputations if the number of imputations is 10 or 100.

Key Words: Missing data, MNAR, multiple imputation, bias, type I error, power

1. Introduction

In longitudinal clinical trials where subjects are treated over a period of time, dropouts may happen during the follow-up period due to different reasons, e.g., adverse events, lack of efficacy, protocol violation, lost to follow-up, etc. Missing data are commonly classified into three types, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Molenberghs and Kenward, 2007). MCAR assumes that the probability of missingness does not depend on either observed data or missing data. MAR assumes that the probability of missingness depend on observed data but not the missing data, which implies that the probability of a subject dropping out is conditionally independent of current and future observations. When missing data are not MNAR or MAR, they are MNAR. If the probability of missingness depends on the missing data, missing data are MNAR.

Statistical methods of handling missing data are valid under the certain assumption on missing data. The mixed-model repeated measures (MMRM) method is valid under MCAR or MAR assumption for continuous endpoints. However, such an assumption might be difficult to verify in practice. Therefore, a sensitivity analysis might be conducted to assess the impact of missing data on the bias of estimation of treatment effect and inflation of type I error rate.

Commonly used sensitivity analysis is based on imputation approach. Last observation carried forward (LOCF) and baseline observation carried forward (BOCF) are the simplest methods, in which missing values are imputed as the last observed values and baseline observations. However, these two methods are criticized owing to the poor performance in the estimation of treatment effect. In contrast to single imputation, multiple imputation approach creates a collection of complete data sets. Each data set is then analyzed separately, and the collection of analyses is combined via Rubin's approach (Little and Rubin, 2002). In this paper, we compare several multiple imputation methods when data are potentially MNAR.

The rest of the paper is organized as follows. We present four methods to handle missing data under MNAR assumption in Section 2. We then conduct a simulation study to compare the performance of different methods in Section 3. Section 4 concludes the paper with summaries and discussions.

2. Imputation Methods

We introduce four multiple imputation methods which are available widely in the literature.

Group Mean (GM): This approach will have, at a given visit, random draws generated from the normal distribution with the mean and standard deviation of the observed values and have missing values replaced by the random draws, in the respective treatment arms.

For example, for subject *i*, assume (x_{i1}, …, x_{ik}) is an observation vector, if all data observed, in the control arm. At a given visit *j*, compute $\bar{x}_j$ and $\hat{\sigma}_{1j}$ as the observed mean and standard deviation for the placebo arm and impute a missing observation by random draw from $N(\bar{x}_j, \hat{\sigma}_{1j}^2)$. Similar approach is used to impute missing observations in the drug arm, except with the observed mean $\bar{y}_j$ and standard deviation $\hat{\sigma}_{2j}$ computed from the drug arm.

This method is anti-conservative in general but selected to explore the degree of impact on the statistical inference.

Jumping to Control (JC): At a given visit, missing data are imputed using random draws generated from a normal distribution with the mean and standard deviation equal to the observed mean and standard deviation from the placebo arm at that visit. This approach imputes missing data in subjects on the drug arm under the assumption subjects who stop taking the drug will no longer benefit from it in the future, and tend to have outcomes similar to those in the control arm.

Mathematically, at a given visit *j*, all missing observations, either in the placebo arm or in the drug arm, are imputed by random draws from $N(\bar{x}_j, \hat{\sigma}_{1j}^2)$. Note that $\bar{x}_j$ and $\hat{\sigma}_{1j}$ are the observed mean and standard deviation from the placebo arm.

Copying Difference from Control (CDC): At a given visit, the observed mean difference between this visit and the previous visit in the placebo arm and the standard deviation of the differences are computed. Then random draws are generated from the normal distribution based on the observed mean and standard deviation as the differences of

between the observed value from previous visit and missing value at the current visit. The missing values are imputed by adding the differences to the observed values.

At a given visit j, compute the observed difference $d_{ij}=x_{ij}-x_{i,j-1}$ in the placebo arm and let $\bar{d}_j$ and $\hat{\tau}_{1j}$ be the mean and standard deviation of these differences. A missing observation in either drug group or placebo group is imputed by the sum of the subject's observation at visit j-1 and a random draw from $N(\bar{d}_j, \hat{\tau}_{1j}^2)$.

Last Z-Score Carried Forward (LZCF): At a given visit, the z-score of observation from the previous visit is computed for each subject. The missing observation is imputed by using the calculated z-score from the previous visit. At a given visit j, let $\bar{x}_{j-1}$ and $\hat{\sigma}_{1j-1}$ be the observed mean and standard deviation in the placebo arm. Let $z_{ij-1} = (x_{ij-1} - \bar{x}_{j-1})/\hat{\sigma}_{1j-1}$ be the z-score for subject i in the placebo arm. Impute a missing x_{ij} by a random draw from $N(\bar{x}_j, (z_{i,j-1}\hat{\sigma}_{1j})^2)$ for the placebo arm, where $\bar{x}_j$ and $\hat{\sigma}_{1j}$ are the observed mean and standard deviation in the placebo arm at visit j. Missing values at all subsequent visits are imputed in the same way. For the drug arm, the same approach is used.

The idea of this imputation is to preserve the subject's relative position before dropout and after dropout within each arm. If a subject is the half standard deviation away from the center of the group (mean) before dropping out, the subject, with imputed value, will remain the half standard deviation away from the center of the group in the period when the subject drops out. This method was proposed by Hendrix and Wilcock (2009).

3. Simulation Study

A simulation study is conducted to compare 4 multiple imputation methods mentioned in the previous section to answer the following questions:

(1) Which method is the best in terms of minimizing bias and type I error control? Is it JC as recommended in some literature?
(2) Is the best method reasonable in terms of minimizing bias and type I error control?
(3) Does the best method have acceptable power?

3.1 Clinical Trial Background

The simulation study is set up to mimic a real clinical trial. Consider a randomized clinical trial comparing a test drug to a placebo in subjects with Alzheimer's disease. The co-primary efficacy endpoints of the trial are the changes from baseline in ADAS-Cog (Alzheimer's disease assessment scale – cognitive subscale) and DAD (Disability assessment for dementia). The trial duration is 78 weeks/18 months and there are 6 post-baseline efficacy measurements, 13 weeks apart. The objective of the trial is to show the test drug reduces patient decline in both ADAS-Cog and DAD at Week 78 relative to placebo. The estimand of interest is the difference in means between the test drug and placebo in all randomized subjects assuming all subjects adhere to assigned treatment and complete the trial.

In this type of longitudinal trial with long duration, some subjects will not be able to stay in the trial for the whole study period. So there will be missing data. The MMRM analysis may not provide valid inference for the targeted estimand when missing data are MNAR.

3.2 Generation of Incomplete Data Sets

Complete clinical trial data are generated from *6*-dimensional multivariate distribution for the placebo arm and *6*-dimensional multivariate distribution for the drug arm. The mean vectors are set equal in computing the type I error rate. For the power calculation, the mean vectors are selected such that the power of the study is 90% at 2-sided α=0.05. The covariance matrices in the multivariate distributions are chosen depending on the scenario considered. Each data vector generated represents change scores from baseline in ADAS-Cog over time. A larger change score implies that the subject deteriorates more from the baseline. For simplicity but without loss of generality, we assume there is no covariate in the trial so no covariate data are generated.

We consider 2 type of missing data: MNAR and MAR. We create MNAR data first. Since we perform a modified intention-to-treat analysis in which subjects are required to have at least one post-baseline visit data to be included in the analysis. So we assume there is no missing data at Visit 1. Also we assume missing data are monotone missing. Without loss of generality, we assume the subject who deteriorates more is more likely to drop out as MNAR, mirroring the reality that a subject is more likely to leave the trial with worse condition. So the probability of dropping out is a monotone function of missing observation. For visit j=2, the probability that observation x_{i2} drops out in the placebo arm is $\Phi((x_{i2}-\mu_{12}/\sigma_{12}+c_{p12})$, where $\Phi(\cdot)$ is the cdf of a standard normal distribution, μ_{12} is the 2nd element of mean vector, σ_{12} is the 2nd diagonal element of covariance matrix, and c_{p12} is a scaling constant such that the proportion of missing data at Visit 2 is equal to a predefined proportion p_{12}, and c_{p12} can be computed via $E[Z + c_{p12}] = p_{12}$ where Z is the standard normal variable. After MNAR data are simulated for visit j=2, MNAR data can be created for visit j=3 with c_{p13} being computed by $E[Z + c_{p13}] = p_{13}/(1 - p_{12})$ so that the missing data proportion is at the target p_{13}. Similarly MNAR data can be created in sequel for visit j=4, 5, and 6, with targeted missing data proportions p_{14}, p_{15} and p_{16} at Visits 4, 5, and 6, respectively. After MNAR data are created, we create MAR data. For MAR data, the probability of being missing is conditionally independent of current and future observations. We simply assume the missing indicator is Bernoulli $(q_{12}/(1 - p_{12} - p_{13} - p_{14} - p_{15} - p_{16}))$ at Visit 2, where q_{12} is the target probability of MAR missing data. Similarly MAR data can be created in sequel for visit j=3, 4, 5 and 6, with missing data proportions being q_{13}, q_{14}, q_{15} and q_{16} respectively.

Missing data in the drug arm can be created similarly when missing data proportions are chosen as p_{2i}'s and q_{2i}'s.

3.3 Analysis of Incomplete Data Set

For each incomplete data set created in Section 3.2, we can impute the missing data and perform the analysis. The MAR data are imputed first using the regression approach

provided by SAS PROC MI. The MNAR data are then imputed using the approaches introduced in Section 2. We then get a complete data set. When we fit MMRM for each imputed dataset, we will get $\hat{Q}_l$ and $\widehat{W}_l$, the point and variance estimates for the treatment difference between the two treatment arms from imputed dataset. Per the method described in Little & Rubin (2002), the combined treatment difference from *m* imputations can be calculated as

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_l \ .$$

The within-imputation variance is

$$\bar{W} = \frac{1}{m}\sum_{l=1}^{m} \widehat{W}_l \ ,$$

and the between-imputation variance is

$$B = \frac{1}{m-1}\sum_{l=1}^{m} (\hat{Q}_l - \bar{Q})^2 \ .$$

Then the total variance of $\bar{Q}$ based upon *m* imputed datasets is

$$T = \bar{W} + \left(1 + \frac{1}{m}\right) B \ ,$$

whose associated degrees of freedom for *t* distribution is

$$\nu_m = (m-1)\left[1 + \frac{\bar{W}}{(1+m^{-1})B}\right]^2 \ ,$$

and the adjusted degrees of freedom is

$$\nu_m^* = \left[\frac{1}{\nu_m} + \frac{1}{\hat{\nu}_{\text{obs}}}\right]^{-1} \ ,$$

where $\nu_{\text{obs}} = (1-\gamma)\nu_0(\nu_0+1)/(\nu_0+3)$, $\gamma = (1+m^{-1})B/T$, and ν_0 is the complete-data degrees of freedom.

The *p*-value is computed from distribution $t_{\nu_m^*}(\bar{Q}, T)$. A claim that the drug is significantly different from placebo can be made if the *p*-value is no larger than 0.05.

3.4 Simulation Results

To compare different imputation approaches, we can repeat steps stated in Section 3.2 and Section 3.3 many times. The simulated power or type I error rate can be computed as the number of times where the p-value ≤ 0.05 divided by the number of repetitions. The bias of estimate of treatment difference can be calculated as the average of treatment difference between the observed and true one over all repetitions. To provide a better reference, the bias is standardized, i.e., the bias is rescaled by dividing the calculated average using the (true) standard deviation. For power simulations, we used 1000 repetitions and for type I error, we used 5000 repetitions.

There are a few factors which could impact the performance of each imputation method: (1) variance matrices used to simulate the complete data, (2) missing proportions in each arm and distribution of missing data across different visits, (3) number of imputations.

We started with an un-structured variance matrix for both arms. The un-structured matrix was based on data from a clinical trial. We also tried an un-structured variance matrix with its elements produced randomly and the conclusions based on that matrix are the same as presented below.

When all MNAR data occur in the placebo arm (the extreme situation of more missing in the placebo arm), the results favor the placebo arm (less likely to detect the drug beneficial effect). This is because all "bad" values in the placebo arm are missing and imputed largely by "good" values which are based on observed values in the placebo arm. The same conclusion holds when there are more missing values in the placebo arm. We confirmed this via simulations. Because favoring placebo raises little regulatory concern, larger portion of missing in the placebo arm is not of main interest and will not be further investigated.

The impact of MNAR data was evaluated first. Table 1 considers 3 missing data scenarios: (1) There are 15% MNAR data in the drug arm and 0% in the placebo arm, and there is no MAR data; (2) There are 10% MNAR data in the drug arm and 0% in the placebo arm, and there is no MAR data; (3) There are 8% MNAR data in the drug arm and 0% in the placebo arm, and there is no MAR data. Among all imputation methods considered, CDC appears to work best. It has the type I error rate about 10% when MNAR at 10% or below. As noted, JC is not conservative at all, and the type I error rate by JC is at least twice of the rate by CDC. Also the bias by CDC also appears as the smallest. So we will focus on CDC in the following.

Table 1: Simulation Results When There Are More MNAR Data in the Drug Arm Than in the Placebo Arm and No MAR Data

	MNAR = D15%, P0% MAR= 0%			MNAR = D10%, P0% MAR= 0%			MNAR = D8%, P0% MAR= 0%		
	Power	Type-I Error	Bias	Power	Type-I Error	Bias	Power	Type-I Error	Bias
Complete	88.4	5.3	-.001	90.7	5.0	.008	90.2	5.4	-.000
JC	99.5	37.1	.150	99.2	22.9	.116	98.9	16.5	.095
CDC	96.6	11.7	.073	96.2	8.2	.054	95.5	6.9	.043
GM	99.8	50.6	.177	99.7	30.2	.130	99.3	20.4	.103
LZCF	99.7	46.7	.177	99.6	27.7	.130	99.3	18.9	.103

Missing data are distributed equally across visits. The number of imputations =10.

At 10% difference of MNAR between the two treatment arms, more scenarios are explored to study the impact of different percentages of missing data. The simulation results are presented in Table 2. The general trend in this regard is that smaller ratios between the two arms (25%/15% < 20%/10% < 15%/5%) lead to smaller inflation of type I error rate, which is about 6%. This makes sense as when the difference between

arms is fixed and a small ratio indicates that there is no severe off-balance in missing data between 2 arms. An important implication of this observation is that as long as the difference of the MNAR proportions between 2 arms is under good control (say, less than 10%), larger percentage of overall MNAR may not impact results more negatively than smaller overall MNAR percentage.

Table 2: Simulations Results When the Difference of Proportion of MNAR Data between 2 Arms Is 10%

	MNAR = D15%, P5% MAR=0%			MNAR = D20%, P10% MAR=0%			MNAR = D25%, P15% MAR=0%		
	Power	Type-I Error	Bias	Power	Type-I Error	Bias	Power	Type-I Error	Bias
Complete	89.6	5.2	-.002	91.4	4.8	-.003	90.2	5.0	.001
JC	98.1	14.2	.091	97.6	9.7	.077	94.3	7.9	.079
CDC	94.7	6.2	.045	95.4	5.4	.040	91.7	4.6	.045
GM	99.3	24.1	.107	99.6	19.7	.096	99.0	22.2	.105
LZCF	98.7	19.5	.107	99.5	14.6	.096	98.1	16.5	.105

Missing data are distributed equally across visits. The number of imputations =10.

The presence of MAR in the data does not appear to have any impact on the conclusions drawn early on CDC. Regardless how the MAR distributes between 2 arms, the results as presented in Table 3 are generally identical across all scenarios. This conclusion holds true also for larger percentage of MAR (20% instead of 5%).

Table 3: Simulation Results When MAR Missing Data Are Added

	MNAR = D15%, P5% MAR=D5%, P0%			MNAR = D15%, P5% MAR=D2.5%, P2.5%			MNAR = D15%, P5% MAR=D0%, P5%		
	Power	Type-I Error	Bias	Power	Type-I Error	Bias	Power	Type-I Error	Bias
Complete	92.2	5.2	-.000	91.6	5.3	.000	90.1	5.3	.002
JC	98.2	13.9	.092	98.7	14.2	.093	97.7	14.3	.095
CDC	95.5	6.7	.046	94.7	6.3	.047	93.9	6.8	.050
GM	99.6	23.2	.108	99.6	23.9	.109	99.3	23.9	.112
LZCF	99.3	18.7	.109	98.7	19.2	.109	98.7	19.3	.112

Missing data are distributed equally across visits. The number of imputations =10.

In all previous simulation results, we did not discuss the power performance of CDC. As one can see, the power of CDC is very reasonable.

Next we investigate whether the missing pattern will impact the conclusions we have drawn. In table 4, “Even” means that the missing data are distributed evenly across visits in both arms so the results are identical to the results in Table 3. “LPLD” means that the missing data occur more at the later visits in both placebo arm and drug arm. “LPED” means that the missing data occur more at the later visits in the placebo arm but more at the earlier visits in the drug arm. “EPLD” means that the missing data occur more at the earlier visits in the placebo arm but more at the later visits in the drug arm. “EPED”

means that the missing data occur more at the earlier visits in both arms. Per Table 4, the type I error rate by CDC is robust to missing data pattern, while JC varies a lot pending on different missing data pattern.

Table 4: Type I Error Rates with Different Missing Data Patterns

	Missing Data Pattern				
MNAR=D 25%, P 15%; MAR=0%	Even	LPLD	LPED	EPLD	EPED
Complete	5.0	5.5	5.3	5.4	5.2
JC	7.9	8.3	1.2	24.9	4.2
CDC	4.6	4.3	6.3	4.0	5.5
GM	22.2	22.7	5.7	49.4	14.5
LZCF	16.5	16.5	4.0	42.9	12.1

The number of imputations =10.

We also investigate the need to increase the number of imputations from 10 to 100. Per Table 5, it appears that 10 is sufficient.

Table 5: Type I Error Rate with Different Imputation Numbers

MNAR=D 15%, C 0%; MAR=0%	100 Imputations	10 Imputations
Complete	5.3	5.3
JC	37.4	37.1
CDC	11.4	11.7
GM	50.8	50.6
LZCF	47.4	46.7

Missing data are distributed equally across visits.

3.5 Why JC Is Less Conservative Than CDC

From simulation results, we see that CDC is more conservative than JC. Is it true all the time or does that just happen to be true for the simulation scenarios considered? In fact, CDC is more conservative than JC if (1) subjects who are getting worse are likely to drop out and (2) there are more missing data in the drug arm than in the control arm.

Since subjects who are getting worse are likely to drop out, the missing values are likely worse than the average value in the placebo arm. Replacing missing values by the mean in the placebo arm is actually replacing worse values by a less bad value. As there are more missing data in the drug arm, there are more replacements in the drug group. So JC is less conservative. For CDC, the missing values are replaced by the sums of the last observed values, which are worse than the average value of placebo arm at the same visit, and difference in means in the placebo arm, so the replaced values are not as good as the mean in the placebo arm.

4. Conclusions and Discussions

MNAR data impose unique challenges in data analysis. Because the missing values are neither observable nor depend on observed values, there is no easy remedy to correct the impact of missing data on the data analysis. This simulation study has investigated four imputation methods and identified CDC as the top choice in terms of protecting type I error rate. CDC controls the type-I error rate reasonably well (~6.0%) when (1) the difference of percentages of MNAR data between 2 groups <10%, (2) the ratio of percentages of MNAR data between 2 groups <3, and/or (3) there exist MAR data. We also concluded that the number of imputations =10 is good enough.

This paper draws the conclusions mainly based on simulation results. To make the conclusions are reliable, we have done extensive simulations to cover all different scenarios. The conclusions appear consistent across all scenarios we considered.

Acknowledgements

We thank Dr. Lijia Wang for helping on some simulations.

References

Molenberghs G and Kenward MG (2007). *Missing Data in Clinical Studies.* John Wiley & Sons, Ltd.

Little RJA and Rubin D (2002). *Statistical Analysis with Missing Data* (2nd Edition). Wiley-Interscience.

Roger J (2011). Sensitivity analyses that address missing data issues in longitudinal studies for regulatory submission. ASA BioPharm Webinar Slides.

Hendrix SB and Wilcock GK (2009). What we have learned from Myriad trials. J Nutr Health Aging. 13(4):362-4.