MI FOR MI, OR HOW TO HANDLE MISSING INFORMATION WITH MULTIPLE IMPUTATION

Daria Shamaida, Quartesian Clinical Research, Kyiv, Ukraine

ABSTRACT

Missing data is the typical problem in almost all clinical trials studies, that could easily undermine the reliability of the results of clinical research. If missing data is inevitable, it is necessary to learn how to handle the problem appropriately. In some cases, it is possible to ignore missing data while using only available non-missing values. But if there is no sufficient reason for omitting missings...? In this case, the multiple imputation (MI) could help! It is one of the commonly used methods for dealing with missing information in clinical research.

This paper reviews different types and approaches of using MI and describes the benefits of using this tool in both primary and sensitivity efficacy analysis. The implementation of SAS procedures for creating multiple imputations is carried out to determine the cases of using different methods of imputation. Additionally, it presents the analysis and compares the results of imputations.

INTRODUCTION

Uncomplete clinical trial data is typical for most of the studies in the industry. The missingness have a big influence on the results of the analysis because it makes it difficult to perform statistical analysis. Missing data introduce the elements of bias and invalidate the conclusions. There are different methods of dealing with missing information – from using only complete dataset (approach is called complete case analysis) to replacing missingness with imputed values. Most SAS procedures remove missing information by default. This approach could result in unbiased estimates if the rate of missing data is quite small – 1-5% and missing values are assumed to be missing completely at random.

But the nature of missing cases could carry additional information, so the exclusion of subjects with missing values could lead to underestimation of variance and incorrect results. Different approaches to replacing missing values could be performed. One of the most common is multiple imputation, the main idea of that is to produce M datasets imputed with different values (according to the model specified), analyze all imputed datasets separately, and then to combine the estimates obtained from these M datasets.

MISSING DATA PATTERNS

Missing information could take many forms and can be caused by various reasons. To implement the appropriate strategy of dealing with missing data, it is necessary to identify the particular distribution of missing observation (also called a pattern). There are two the most common missing data patterns:

- Monotone missing data pattern.
- Arbitrary pattern of missing data (also called non-monotone).

The first one occurs when the value of the variable is missed not only in the specific time point for the particular subject, but in all subsequent time occasions for this individual. The common example of such observation distribution is when a subject is lost for a follow-up period but have the treatment phase completed.

The general scheme for monotone missing data pattern is presented below:

Obs	Variable			
	V1	V2	V3	V4
1	X	X	x	x
2	X	X	x	o
3	X	x	o	o
4	X	o	o	o
5	O	o	o	o

Table 1 Monotone missing data pattern

The monotone pattern of missing information lends itself to approaches that have simpler assumptions than methods required for an arbitrary pattern of missing data.

The non-monotone missing data pattern indicates a situation when there is no particular pattern in the missing data structure (the missing observations are distributed in a nonsystematic way across different cases and variables). The example, if the record is missed for a subject at a given time point, but returns later at the follow-up visit.

An example of a non-monotone pattern is given below:

Obs	Variable			
	V1	V2	V3	V4
1	X	o	o	x
2	X	x	o	x
3	O	x	x	o
4	X	o	o	x
5	O	x	x	x

Table 2 Non-monotone missing data pattern

In addition, there are many other missing data patterns. One of them is missing by a design pattern (also called "matrix sampling pattern"). This type of missing data pattern arises in studies with randomization procedures to allow item missing data on selected variables for subsets of study observations. The general example of this pattern is given in table 3:

Obs	Variable			
	V1	V2	V3	V4
1	X	x	o	x
2	X	x	o	x
3	X	o	x	o
4	X	o	x	o

Table 3 Missing by design pattern

So, the missing data pattern depends on the nature of data missings. And the subsequent strategy and method used are determined by missing data pattern and type of variables.

MISSING DATA MECHANISM

Before specifying the definite methods to be used for the analyses, it is important to define the type of missingness. Missing information for the single variable refers to one of the following categories: MCAR (missing completely at random), MAR (missing at random), and MNAR (missing not at random).

The missing data are **MCAR** if the probability that information is missing does not depend on observed or unobserved data. Under MCAR the missing values are the simple random sample from all values, and there are not systematic rules of facing the missing data values. This mechanism states that missingness is not related to any factor, known or unknown, in the study. For example, such a situation could occur when an automatic sphygmomanometer is broken down, and some patients` blood pressure measures were not recorded.

If the missing data depends on the observed responses but is not related to the missing values, the mechanism of missing data is **MAR**. In practice, the MAR is the most common mechanism used for multiple imputations. Under MAR, the missing-data values do not contain any additional information given observed data about the missing-data mechanism., thus the process that causes missing data can be ignored. For example, the depression survey data is MAR, given gender, because men are less likely to participate in different surveys. Once the gender is counted, then the missing values do not depend on the level of their depression.

The missing data is missing not as random (**MNAR**) if the systematic difference after taking into account remains between observed values and missing values. For example, subjects with severe adverse effects are less likely to complete the depression survey. Thus, the reason of missingness is related to the unobserved characteristic of the subject

STRATAGIES OF DEALING WITH MISSING DATA

There are plenty of methods that could be applied to the missingness depending on the goals of research and the features of data studied. The most common are: complete case analysis, complete case analysis with weighting adjustments, full information maximum likelihood, expectation-maximization algorithm, single imputation of missing values, and multiple imputation (MI).

The simplest approach of handling missingness is **complete cases analysis**, which involves the list-wise deleting of missing values. Most SAS® procedures use list-wise removal by default. The estimates will be unbiased only when data is missed under MCAR. When data is missed under MAR, the complete case analysis is applicable only if the item missing data rates from 1% to 5% - in this case, not much information will be lost and the bias will not a major. The usage of this approach is not efficient when the missingness rate increases. For example, if analysis data has the missing patterns presented in Table 1 or Table 2, only the first and fifth patterns respectively will be analyzed (since they have no missingness across all variables). Hence a lot of information will be lost and the estimates will be biased and the accuracy of results supposed to below.

The advantages of using this method are:

- Simplicity – no additional data modifications are required since the complete data statistical analysis could be used.
- Comparability of univariate statistics – since the sample base of cases is common.

Disadvantages stem from the loss of information in deleting incomplete cases, this results in loss of accuracy and biased estimates when missed data is not MCAR.

The second method of dealing with missing data is a **complete case analysis with weighting adjustments**, which is a modified version of the previous approach. Usually, the usage of weighting is limited to monotonic missing patterns, when a large number of variables are missed for each case. The idea of this approach is to analyze complete cases by introducing additional to compensate for an item missing data on one or several variables. In practice, the usage of this method could cause several difficulties, because different adjustment factors would be needed for each target variable. Also, weight-by-variable strategy would work for univariate analysis but would be a problem to chose the variable-specific weights for the multivariate analysis. The next difficulty is that access for data that used for base weighting adjustments is restricted, which does not make the later weighting possible.

The strategy for dealing with some item missing data problems is several maximum likelihood methods, such as full information maximum likelihood and expectation-maximization algorithms. These methods are used to find the estimates of parameters ($\hat{\theta}$) that maximize the likelihood $L(\hat{\theta}|Y_{obs})$. **Full information maximum likelihood (FIML)** methods directly maximize the likelihood function in which incomplete cases support only to the estimation of parameters for which the sufficient statistics are functions of the observed values for the case (e.g., μ1,σ1 for observed Y1; σ12 for observed {Y1, Y2}). The FILM methods require the parametric likelihood for data defined by the user, which could cause some difficulties in complex sample survey data where weighting is needed and informative stratification and clustering of observational units makes it difficult to determine the real form of the data likelihood. Structural equation modeling (SEM) or other latent variable model analyses are the applications for FILM, in which Y is assumed to follow a multivariate normal

distribution with parameter vector θ={μ, Σ}. The SAS procedure PROC CALIS is a typical tool to perform the SEM or other FILM methods.

The next maximum likelihood method that could be used for generating maximum likelihood estimates is the **expectation-maximization (EM) approach**. This algorithm requires the parametric likelihood function for complete data, the same as the FILM approach. The idea of this method is that there are two steps, that are performed iteratively – expectation and maximization to derive the maximum likelihood parameter estimates. The expectation (E) step replaces the missingness with their expected values under the current iteration estimates and constructs the corresponding sufficient statistics. Then the maximization (M) step maximizes the complete data likelihood to obtain the new estimates of parameters of the model. These two steps repeat until the parameters converge. Assuming the complete data for analysis is distributed as multivariate normal, MVN(μ,Σ) the PROC MI with EM statement could be used to compute maximum likelihood estimates of the mean and covariance parameters. Using the OUTEM= option of the EM statement in PROC MI the estimated mean vector and covariance matrix can be output to a designated file, defined by the user. The "completed" dataset with missingness replaced by the EM estimates could also be output. But the "imputed" dataset generated in such a way does not account for two sources of variability:

1) Residual variance - variability of the true values about their expectations.
2) The imputation variability that is inherent in the estimation of the expected values.

These two sources of variability are reflected in a proper MI treatment of the missing data problem. The EM is very important because the "completed" dataset generated by the EM algorithm is used as the matrix of starting values for the iterative MCMC multiple imputation procedure.

The next tool for handling missing data is the **single imputation methods**. It involves the usage of a single imputed value replacing missing data on key variables. In general, imputations are means or draws from a predictive distribution of the missing values. They require a method of creating the predictive distribution for the imputation based on observed data. There are two common classes of methods to generate this distribution: explicit and implicit.

Explicit modeling means that predictive distribution is based on a formal statistical model, and hence the assumptions are explicit. The common explicit approaches are mean imputation, regression imputation, and stochastic regression imputation. In mean imputation means from responding units in the sample are substituted. The means could be formed within classes or cells. The idea of regression imputation is to replace missing values by predicted values from a regression of the missing item on items observed for the unit, usually calculated from units with both observed and missing variables present. Mean imputation could be considered as a special case of regression imputation where the predictor variables are dummy indicator variables for the cells within which the means are imputed. By using stochastic regression imputation missing values are replaced with the value predicted by regression imputation plus a residual, drawn to reflect uncertainty in the predicted value. In normal linear regression models, the residuals will be normal distributed with zero mean and variance equal to the residual variance in the regression. With a binary outcome, as in logistic regression, the predicted value is a probability of 1 versus 0, and the imputed value is a 1 or 0 drawn with that probability.

Another approach used in single imputation is *implicit* modeling, where the main focus is on an algorithm, which implies an underlying model. The most popular methods are cold deck imputation, hot deck imputation, substitution, composite methods, and others. The hot deck method involves substituting individual values drawn from ‘’similar" responding units. In the cold deck imputation, the missingness is replaced by a constant value from an external source (for example, the previous realization of the same study). There is also a substitution method, which replaces the nonresponding units with alternative units that were not included in the analysis sample.

The single imputation method is quite simple and the advantage of this is the following: if the imputation technique is multivariate and retains the stochastic properties in the observed data, a single imputation may address potential bias for MAR missing data. But the main disadvantage of the model is that the standard analysis of a singly imputed data set is that it precludes estimation and inference that fully reflects the variance attributable to the item missing data imputations.

And last robust, flexible option for dealing with missingness **is multiple imputation** (MI). A lot of different missing data patterns could be analyzed using this framework with the implementation of various analytic models. The PROC MI SAS procedure is the main tool to perform the multiple imputation: many different approaches and options are available there, and a lot of them reflect the historical aspect of PROC MI evolution. Some approaches were developed for specific missing data patterns (for example, the MONOTONE option), while others assume specific distributions for the variables (e.g., MCMC for multivariate normal data).

WHY THE MULTIPLE IMPUTATION METHOD?

1) Multiple imputation is *model based*. The term imputation model should be defined wider than the model for analysis. Usually, the imputation model is based on explicit distribution, but using implicit methods (for example, the nearest neighbor imputation method) could also have good results. Having the model as a basis ensures statistical transparency and integrity of the imputation process.
2) Multiple imputation is *stochastic*. It imputes missing values based on draws of the model parameters and error terms from the predictive distribution of the missing data, Ymis.
3) Multiple imputation is *multivariate*. This means that not only the observed distributional properties of every single variable are preserved, but also the relationship between variables included in the imputation model may be presented.
4) Multiple imputation *uses multiple independent repetitions of the imputation* that makes it possible to estimate the variability of parameter estimates that are attributable to imputing missing values. This is the variability that is in addition to the recognized variability in the underlying data and the variability due to sampling.
5) Multiple imputation is *robust against not major deviations from theoretical assumptions.* There is no ideal model that could exactly match the random variable distribution or fit the assumed data missing mechanism, but using MI in such cases will obtain robust and valid estimates.
6) Multiple imputation is very *usable* in different statistical software systems (SAS, STAT, R studio, etc.), which makes the applying of this method quite simple and convenient for the user.

But this approach of dealing with missingness is also criticized due to using the imputation model that could differ a lot from the model will be used for the analysis (for example, the important variable of analysis may not be included in the

imputation model, or the interactions between variables could not be taken into the consideration while performing imputation, etc.)

STEPS OF PERFORMING MULTIPLE IMPUTATION

Multiple imputation is not just a technique for imputing data, it is also the tool for obtaining estimates and correct inferences for statistics ranging from simple descriptive statistics to the parameters of complex multivariate models. In general approach, there are 3 steps of performing MI: imputation of M datasets, based on the particular model; analysis of the M imputed datasets; combining the results from step 2 to obtain MI estimation.

But before starting the 1st step it could be useful to perform 0 step – to examine the missing data pattern and group means. In practice, this could be done using SAS procedure PROC MI with NIMPUTE=0 and a SIMPLE option before imputing values. The SIMPLE option is used for displaying univariate statistics and correlations and NIMPUTE=0 means that 0 imputations will be performed.

Step 1 defines the variables and distributional assumptions of the imputation model and applies specific MI algorithms to generate the multiple imputations of the missing values and output a completed data set for each of m=1,…,M repetitions of the imputation process. So this step could be divided into the following 3 sub-steps: specifying the imputation model with defining the data distribution, then impute the missing values, and produce multiple imputations.

Step 2 is to estimate the parameters of interest from M imputed datasets using BY _IMPUTATION_ statement. The model defined is the same that would be used for complete data. The results obtained from each imputation will be different because the imputes data differ. The outputted data from every imputation will be combined to one estimate in the next step.

Step 3 required the usage of PROC MIANALYZE SAS procedure, which inputs the parameter estimates and standard errors from the preceding analysis step and applies the multiple imputation formulae to generate the MI estimate, standard errors, confidence intervals, and hypothesis test statistics for making inferences for descriptive statistics or model parameters.Under appropriate conditions, the pooled estimates are unbiased and have the correct statistical properties.

Hence the detailed scheme of performing multiple imputation is the following:

Step 0. Examine the missing data pattern and group means.

Step 1a. Specify the Imputation model.

Step 1b. Impute the missing values.

Step 1c. Produce Multiple Imputations.

Step 2. Analyze M imputed datasets.

Step 3. MI Estimation and Inference. Combining Step 2 Analysis.

The typical difficulty for this method is to define all variables to be included in the imputation model. In practice, the set of variables for MI should be broader than the list of variables used for the analysis.

It is not possible to include every variable in the imputation model, but according to the recommendations from Schafer (1999) and van Buuren (2012) the logic of choosing variables should be the following:

1. Include all key analysis variables.
2. Include other variables that are correlated or associated with the analysis variables.
3. Include variables that predict item missing data on the analytic variables.

Failure to include one or more analysis variables in the imputation model could result in biased MI estimations. and inference. Including additional variables, that are good predictors of the analytic variables improves the precision and accuracy of the imputation of item missing data. Under the assumption that item missing data is MAR, incorporating variables that are correlated with the variables that have missing data and predict the propensity for response will reduce bias associated with the item missing data mechanism.

Depending on the nature of the variables used as set for the imputation and considering the type of missing data pattern different methods could be used in SAS procedure PROC MI. The general scheme of choosing the appropriate imputation method is presented in table 4:

Missing Data Pattern	Variable Type	MI Method
Monotone	Continuous	Linear regression, predictive mean matching, propensity score
	Binary/ ordinary	Logistic regression
	Nominal	Discriminant function
Arbitary	Continuous	With continuous covariates: MCMC monotone method MCMC full-data imputation
	Continuous	With mixed covariates: FCS regression FCS predictive mean matching
	Binary/ ordinary	FCS logistic regression
	Nominal	FCS discriminant function

Table 4. Imputation Methods

In general, the algorithms used for monotone missing data patterns are simpler than approaches for arbitrary missing data pattern.

The Fully Conditional Specification (FCS) method is a newer approach that is available starting in SAS/STAT 12.1 software with using FCS statement, that dictates how the variables with an arbitrary missing data pattern will be imputed (FCS REG or FCS LOGITSTIC, etc.).

But how to implement this method in the earliest version of SAS? It is needed to break the imputation step into two parts. The first part imputes just enough missing values to create a monotone missing data pattern. It could be done by using the MCMC method with IMPUTE=MONOTONE option with only continuous covariates in the imputation model. (If categorical variables should be taken into account they can be added to a BY statement). The second part is to use the MONOTONE statement to impute the remaining missing data, using the next calling of PROC MI. But the results obtained

from specifying FCS statement and using 2-step approach will differ a bit due to the random nature of how the data are drawn for the imputations

SAS IMPLEMENTATION

A random dummy dataset was generated for the purposes of analysis. The parameter to estimate is Pain Numeric Rating Scale (P-NRS), filled by subjects daily. Weekly Pain NRS values are calculated as the average of all measures during the week. All patients will be treated by either Placebo or Active Drug. The severity of the disease is Mild-Moderate or Moderate-Severe.

Primary Endpoint: Change form baseline in Pain Numeric Rating Scale at Week 5, supposing that missing data is MAR.

Sensitivity Analysis I of Primary Endpoint: Change form baseline in Pain Numeric Rating Scale at Week 5, supposing the missingness are MNAR (based on the reason for withdrawal).

Primary and sensitivity analysis will be performed using mixed effects model with repeated measures (MMRM). The model will contain treatment, week, and treatment-by-week interaction as fixed effects, baseline score, and disease severity as covariates.

Also, the weekly P-NRS score will be set to missing after taking any rescue medication.

Subject ID	TRTP	Severity	Baseline	Week 1	Week 2	Week 3	Week 4	Week 5
001	Placebo	Mild-Moderate	X	X	X	X	O	X
002	Placebo	Moderate-Severe	X	X	X	X	X	O
003	Active	Mild-Moderate	X	X	O	X	O	O
004	Placebo	Mild-Moderate	X	X	X	X	X	O
005	Active	Mild-Moderate	X	X	O	X	X	X
006	Active	Moderate-Severe	X	X	O	O	O	X
007	Placebo	Mild-Moderate	X	X	X	X	O	X

Table 5. Input dataset structure

Step 0. At first, it will be useful to examine the missing data pattern. This could be done by using zero imputation (specifying NIMPUTE=0), also the SIMPLE option will display univariate statistics and correlation. The following code is used for this:

```
proc mi data=for_imputation nimpute=0 out= M_3 simple;
var Week_1 - Week_5;
run;
```

Missing Data Patterns												
Group	Week 1	Week 2	Week 3	Week 4	Week 5	Freq	Percent	Group Means				
								Week 1	Week 2	Week 3	Week 4	Week 5
1	X	X	X	X	X	70	84.34	7.0889	6.2320	5.8740	5.7192	5.3765
2	X	X	X	X	.	3	3.61	7.4761	6.2857	6.5158	5.7555	
3	X	X	X	.	.	2	2.41	4.6428	4.4047	4.8333		
4	X	X	.	X	X	1	1.20	7.4285	8.6000		7.6000	6.6666
5	X	X	.	.	.	1	1.20	5.8571	5.0000			
6	X	.	X	X	X	1	1.20	8.4000		7.5714	7.0000	7.60000
7	X	.	.	.	.	4	4.82	7.2714				
8	O	O	O	O	O	1	1.20					

Table 6 Missing Data Patterns

Step 1. This step is used for specifying the model of imputation and performing the imputations of M datasets.

Missing at random (MAR) – example

Since table 6 indicates that the missing pattern is non-monotone, it is necessary to perform the partial imputation to get the monotone missing data pattern, since the primary endpoint assumes that the data is missed at random (MAR). The Markov Chain Monte Carlo (MCMC) method will be used for implementing multiple imputation. This method is appropriate for non-monotonic missing data. Using the following code the intermittent missing data will be imputed:

```
proc mi data=for_imputation nimpute=10 out= imput_1;
      mcmc impute=monotone ;
      var Severity Baseline Week_1 - Week_5;
      by TRTP;
run;
```

The next step is to perform multiple imputation for monotone missing pattern data using regression:

```
%sort(imput_1,_imputation_);
proc mi data= imput_1 nimpute=1  out= imput_2;
      by _imputation_;
      class TRTP;
      var TRTP Severity Baseline Week_1 - Week_5;
       monotone reg;
run;
```

After that the data obtained in the previous step should be transposed to get it vertically-orientated:

```
%sort(imput_2, usubjid _imputation_);
proc transpose data=imput_2 out=mar;
       by usubjid _imputation_;
       var baseline WEEK_1 WEEK_2 WEEK_3 WEEK_4 WEEK_5;
run;
```

The derivation type for these records (DTYPE variable) will be defined as “MAR” in the corresponding ADaM.

Missing not at random (MNAR) - example

The Sensitivity Analysis I of Primary Endpoint assumes the missing data is missed not at random, based on the reason for discontinuation. This part of the analysis assumes missing Pain Numeric Rating Scale for subjects who discontinue from the study due to usage of any rescue medication or due to an adverse event is MNAR and for other subjects missing data is considered to be MAR. Missingness for subjects who took rescue medication or withdraw due to an adverse event will be imputed based on baseline distribution of weekly scores for all subjects with the same disease severity status assuming trimmed normal (from 5 to 10).

The data obtained in the previous step (where the MI was assuming data is missed at random) will be used for the sensitivity analysis of the primary endpoint with some transformation.

At first, it is needed to select only the data for patients who did not withdraw early due to an AE or taking rescue medication from the dataset obtained after performing multiple imputation for monotone missing pattern data using regression:

```
%sort(imput_1 imput_2, usubjid);

data not_ae_rm ;
       merge   imput_2(in=a)   adsl(in=b   where=(lowcase(DSCREAS)='adverse   event'   or
RESUSFL='Y') keep=usubjid DSCREAS RESUSFL);
       by usubjid;
       if a=1 and b=0 then output not_ae_rm ;
run;
```

In the code above DSCREAS variable contains information about the reason of discontinuation and the variable RESUSFL shows whether the subject took any rescue medication. Using condition `if a=1 and b=0` allow selecting cases for subjects who didn`t discontinue from the study due to AE or who took any rescue medication. Missing values for them will be considered as MAR.

The next step is to select records for subjects who withdrew early due to an AE or took opioid rescue medication from the dataset obtained after performing the partial imputation to get the monotone missing data pattern:

```
data ae_rm ;
       merge imput_1(in=a) adsl(in=b where=(lowcase(DSCREAS)='adverse event’ or RESUSFL
='Y') keep=usubjid DCSREAS RESUSFL);
       by usubjid;
       if a=1 and b=1;
run;
```

Using the condition `if a=1 and b=1` allow getting records for subjects who took RM or discontinued due to AE.

The next step is to combine two datasets obtained above:

```
data all_for_mnar;
       set ae_rm non_ae_rm;
run;
```

After that it is needed to transpose data to get it vertically-oriented and specify the new variable to be used in proc MI that is equal to baseline values:

```
%sort(all_for_mnar,usubjid _imputation_ Severity TPTP);
proc transpose data= all_for_mnar out=ae_mnar ;
       by usubjid _imputation_ Severity TPTP;
       var Baseline-- Week_5;
run;
data base_ae_rm_mnar;
       set ae_mnar;
       if _NAME="Baseline" then new_base=col1;
run;
```

And the final action in step 1 for sensitivity analysis is to perform MI using monotone regression to impute missingness with values from a truncated normal distribution 5 to 10:

```
%sort(base_ae_rm_mnar , _imputation_);
proc mi data=base_ae_rm_mnar out=mnar_ nimpute=1 min=.5 max=.10;
       by _imputation_;
       class ADDISSEVN;
       var ADDISSEVN new_base;
       monotone reg(new_base=ADDISSEVN);
run;
```

The derivation type for these records (DTYPE variable) will be defined as “MNAR” in the corresponding ADaM.

Step 2. This step includes the analysis of M imputed datasets. The mixed effects model with repeated measures (MMRM) will be used in this analysis. This model contains treatment, visit, and treatment-by-visit interaction as fixed effects, baseline score, and disease severity as covariates. Records with the appropriate derivation type (MAR or MNAR) will be used separately to get the results from the following model:

```
proc mixed data = data_for_analysis;
       by _imputation_;
       class trtp visit usubjid severity;
       model chg = trtp  visit trtp*avisit baseline Severity /ddfm=kr;
       repeated visit/ type = un subject = usubjid;
       lsmeans trtp*visit / pdiff cl;
run;
```

Since the dataset used in the model above have 10 imputations it is needed to perform a separate analysis for each imputation – this could be reached by using the `by _imputation_` statement.

Step 3. This step is to combine the results of the analysis obtained from step 2. This should be done with the help of the PROC MIANALYZE SAS procedure. It should be done for both LS mean estimates and LS mean differences:

```
/*combining results for LS means*/
Proc mianalyze data=lsm;
       by visit trtp;
       modeleffects estimate;
       stderr stderr;
Run;

/*combining results for LS mean differences*/
proc mianalyze data=diff;
    by avisitn trtp _trtp;
    modeleffects estimate;
    stderr stderr;
run;
```

After performing all three steps described above the following results was obtained:

	Results under MAR		**Results under MNAR**	
	Treatment A	Treatment B	Treatment A	Treatment B
LS Mean (SE)	-1.995 (0.327)	-2.452 (0.310)	-1.958 (0.329)	-2.145 (0.310)
95 % CI	(-2.636, -1.354)	(-3.061, -1.843)	(-2.603, -1.312)	(2.754, -1.536)
LS Mean difference A vs B (SE)	0.456 (0.421)	x	0.187 (0.426)	x
95 % CI	(-0.369, 1.283)	x	(-0.647, 1.022)	x

Table 7 Results of combing M imputed datasets

Table 7 shows that results obtained from primary analysis and sensitivity analysis of primary endpoint are quite different, so it is necessary to define the missing data mechanism correctly to obtain robust estimates, based on the nature of missing data in every particular case.

CONCLUSION

The missingness is a typical problem for many clinical trial studies, that affect the estimates that obtained, so the results based on data with missingness should be interpreted with caution. Also, missing data reduces the power of a trial, hence requires the increase of the sample size, but it can not guarantee that the estimates are not biased.

In many cases the missing values are replaced by other values – for this purpose many different tools are used, among them there is a multiple imputation:

- MI is a very useful tool for dealing with missing data.
- This method gives robust parameter estimates.
- It is applicable when other strategies for dealing with missing data could not be applied.
- It uses different methods, depending on the type of missing variable and pattern.
- It allows implementing sensitivity analysis of primary endpoints, assuming another missing data mechanism, or using another imputation method.

REFERENCES

Berglund, Patricia and Heeringa, Steven, 2014. Multiple Imputation of Missing Data Using SAS®. Cary, NC: SAS Institute Inc.

Horton, N.J. and Lipsitz, S.R. (2001), “Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables,” Journal

Kenward MG. The handling of missing data in clinical trials. Clin. Investig. (Lond.) 3, 241–250 (2013).

Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc.

SAS Institute Inc. 2015. SAS/STAT® 14.1 User’s Guide. Cary, NC: SAS Institute Inc.

Schafer, J. L. 1996. *MIX: Multiple Imputation for Mixed Continuous and Categorical Data*. Software library for S-PLUS. Written in S-PLUS and Fortran-77. Available at http://www.stat.psu.edu/~jls/misoftwa.html.

CONTACTS

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Daria Shamaida
Company: Quartesian
Address: 142 Antonovycha street
City / Postcode Kyiv 03150, Ukraine
Work Phone: +380954224771
Email: daria.shamaida@quartesian.com
Web: https://quartesian.com/