PharmaSUG China 2025 - Paper SA-187

Step-by-Step Implementation Procedures of Multiple Imputation for Missing Data

Binhan Chen, Sichen Li, Xinran Tang, InnoCare Pharma Limited

ABSTRACT

Missing data is a common issue in a substantial number of statistical analyses. Multiple imputation (MI) is a technique for handling missing data. Various methods have been developed and are readily available in SAS/STAT® software PROC MI for multiple imputation of both continuous and categorical variables. In the use of PROC MI, different parameters need to be selected according to the specific data situation.

Systemic Lupus Erythematosus (SLE) is a disease in the field of autoimmune disorders. The SLE Responder Index (SRI) serves as a widely adopted method for evaluating SLE efficacy. The SRI incorporates diagnostic criteria from three internationally recognized indices: the SELENA-SLEDAI (Safety of Estrogens in Lupus Erythematosus National Assessment–Systemic Lupus Erythematosus Disease Activity Index), the Physician's Global Assessment (PGA), and the British Isles Lupus Assessment Group (BILAG) 2004 edition.

In clinical projects with missing efficacy endpoint SRI-4 response data, multiple imputation can be implemented through two analytical approaches: The first approach imputes missing SRI-4 values directly; the second approach separately imputes missing data for each of the three component indices (SELENA-SLEDAI, PGA, and BILAG-2004) before calculating the SRI-4. Each method presents distinct advantages. This paper demonstrates SAS implementations for both approaches and discusses differences in analytical outcomes derived from them.

INTRODUCTION

Missing data are data that were planned to be collected in clinical trial but are not collected and recorded in the database. No matter how well designed and conducted a trial is, some missing data can almost always be expected. MI is a useful tool for conducting analyses under both missing at random (MAR) and missing not at random (MNAR) assumptions (Michael O'Kelly, Bohdana Ratitch, 2014, p. 256). Multiple imputation methods generate multiple copies of the original dataset by replacing missing values using an appropriate stochastic model, analyze them as complete sets and finally combine the different parameter estimates across the datasets to produce a unique point estimate and standard error taking into account the uncertainty of the imputation process.

Systemic lupus erythematosus (SLE) is a complicated multisystem autoimmune disease that carries substantial mortality and morbidity. The efficacy endpoint is SLE Responder Index (SRI)-4 response rate. SRI-4 response is defined as: a) $\geqslant$4 points reduction from baseline in SLE Disease Activity Index-2000 (SLEDAI-2K) score; and b) No worsening (increase of <0.3 points from baseline) in Physician's Global Assessment (PGA); and c) No new A organ domain score or no more than 1 new B organ domain scores compared with baseline in British Isles Lupus Assessment Group (BILAG)-2004.

EXAMPLE DATASET

Analysis in this paper will be illustrated using an example dataset with the following variables:

subjid – subject identification number;

grp – group (T=treatment and P= placebo);

factors – the stratification factors;

cycle1, cycle2, cycle3 – binary variables representing response to treatment (Y=responder; N=non-responder) at post-baseline study visits 1, 2, and 3 respectively;

resp_1, resp_2, resp_3 - continuous variables representing response score to treatment at post-baseline study visits 1, 2, and 3 respectively;

base – a continuous baseline score.

STATISTICAL METHODOLOGY

In this paper, stratified Cochran-Mantel-Haenszel (CMH) chi-square test will be used to test the difference between the treatment and the placebo, with specific implementation methodology referenced from Ratitch, B. and O'Kelly, M. (2013).

A simulated dataset of 100 subjects was generated (50 per group), and an adjusted CMH difference rate of 18.6% (-0.6%, 37.8%). Missing data were assumed to follow a missing at random (MAR) mechanism, with an overall 10% missingness rate in efficacy endpoints.

Code

The code below shows how to generates 50 random numbers, and then uses these random numbers to perform random missing data processing on the complete dataset, resulting in 50 different datasets with missing values.

```
data random_numbers;
    call streaminit(123456);
    do id = 1 to 50;
        uniform_random = rand('uniform');
        RandomValue = round(uniform_random,0.00001)*100000;
        output;
    end;
run;
%macro imput(seed=,number=);
  data datain_&number;
    set alldata;
    call streaminit(&seed.);
    array periods {3} cycle1 cycle2 cycle3;
    do i = 1 to dim(periods);
        if rand('uniform') < 0.1 then periods[i] = '';
    end;
  run;
%mend;
data _null_;
    set random_numbers;
    call execute("%imput(seed="||RandomValue||",number="||id||");");
run;
```

In general, the MI is conducted through 3 steps. 1) The missing data are filled in *m* times to generate *m* complete data sets. 2) The *m* complete data sets are analyzed by using standard procedures. 3) The results from the *m* complete data sets are combined for the inference.

SAS procedure PROC MI offers several methods for imputation of both continuous and categorical variables (SAS User's Guide, 2015). The choice of method primarily depends on whether the type of missing data pattern is monotone and the imputed variables type.

Now we will briefly present the SAS code for implementing these two methods.

METHODS 1 - IMPUTES MISSING SRI-4 VALUES DIRECTLY

Step.1. Check the missing patterns.

Code

The code below shows how to use PROC MI to check the missing patterns.

```
proc mi data=datain nimpute=0;
    class cycle1 cycle2 cycle3;
    fcs logistic;
```

```
    var cycle1 cycle2 cycle3;
run;
```

Missing Data Patterns					
Group	cycle1	cycle2	cycle3	Freq	Percent
1	X	X	X	72	72.00
2	X	X	.	5	5.00
3	X	.	X	11	11.00
4	X	.	.	1	1.00
5	.	X	X	10	10.00
6	.	.	X	1	1.00

Figure 1. Missing Data Pattern

We can observe that the dataset is Non-monotone pattern. Since the imputed variables type are binary categorical variables, the Fully Conditional Specifications (FCS) logistic regression method is selected for imputation.

Step.2. Use MI to generate complete datasets.

Code

The code below shows how to do the multiple imputation for SRI-4 values by FCS logistic regression method.

```
proc mi data = datain seed = 123456 nimpute = 10 out = sri_imp;
    class grp factors cycle1 cycle2 cycle3;
    fcs logistic (cycle1 = grp factors);
    fcs logistic (cycle2 = grp factors cycle1);
    fcs logistic (cycle3 = grp factors cycle1 cycle2);
    var grp factors cycle1 cycle2 cycle3;
run;
```

Through this program, we will generate a completed datasets which have 10 copies of the original dataset. The class statement defines all categorical variables encountered in the MI procedure. To minimize the impact on the results obtained from different methods, a seed value of 123456 is used for all imputations in this paper, and the number of imputation iterations is set to 10.

Meanwhile, several issues often arise during the imputation process, such as:

- Time dependency handling: If the project efficacy has time dependence, and the previous efficacy results are used as influencing factors in the program.
- SAS warning due to single-value stratum from insufficient data: Insufficient data volume leads to only a single value in a certain stratum after stratification, resulting in a SAS warning.
- SAS warning due to the maximum likelihood estimates for the FCS method logistic model for variable xxx in an iteration process may not exist: Insufficient data volume has prevented the model from completing its task. The FCS discim method might be able to solve this warning.
- SAS error due to each observation has analysis variables either all missing or all observed in the data set: During the imputation process, if values in a stratum are all missing, SAS will prompt an error, and the program cannot run, assigning a single value may be considered. However, it will inevitably introduce bias and should be used with caution.

Step.3. Analyze the completed datasets using standardized procedures and combine results for the inference.

Code

The code below shows how to use the stratified CMH chi-square test method to combine multiple imputed datasets and generate the risk difference.

```
*** Estimate proportions of responders in each treatment arm;
proc freq data= sri_imp;
 tables avaln / cl binomial(level='1');
 by _imputation_ grpn grp factors;
 ods output binomialprop=prop;
run;


*** From ODS output dataset BINOMIALPROP, create a dataset containing
estimated proportion of responders in each treatment arm and their standard
errors;
data prop_trt;
    merge
    prop(where=(label1="proportion") keep=_imputation_ grp nvalue1 label1
rename=(nvalue1=prop))
    prop(where=(label1="ase") keep=_imputation_ grp nvalue1 label1
rename=(nvalue1=prop_se));
    by _imputation_ grp;
run;


*** Combine proportion estimates;
proc sort data=prop_trt; by grp _imputation_; run;
proc mianalyze data=prop_trt;
    modeleffects prop;
    stderr prop_se;
    by grp;
    ods output parameterestimates=mian_prop_trt&ord.;
run;


*** Compute estimates of the difference in proportions of responders
between treatment arms and their standard errors;
data prop_diff;
 merge prop_trt(where=(grpn= 1) rename=(prop=p1 prop_se=se1))
     prop_trt(where=(grpn=2) rename=(prop=p2 prop_se=se2));
    by _imputation_;
    prop_diff = (p2-p1);
    se_diff = sqrt(se1*se1 + se2*se2);
run;
*** Combine estimates of the proportion differences;
proc mianalyze data=prop_diff;
    modeleffects prop_diff;
    stderr se_diff;
    ods output parameterestimates=mian_prop_diff;
run;
```

The above code is cited from Ratitch, B. and O'Kelly, M. (2013). For a detailed understanding of the derivation process, please refer to the article.

METHODS 2 - SEPARATELY IMPUTES MISSING DATA FOR EACH OF THE THREE COMPONENT INDICES

Since each evaluation item has a different definition and the data types are not the same, different impute methods need to be selected for processing.

The SLEDAI-2K total score and PGA score are continuous variables, while the other is a binary variable.

Since the missing values of SRI-4 in the data simulation process of this paper are consistent with the missing values of the three indicators, both of the missing data pattern are non-monotone. Meanwhile, the method used in analyzed and combined is also consistent with Methods 1 step.3., so it will not be displayed repeatedly.

Therefore, this chapter only shows the specific methods and codes of PROC MI for different types of variables.

Step.1. - continuous variables

The SLEDAI-2K total score and PGA score are numerical variables, and both of which are associated with baseline results, so the baseline value needs to be added as influencing factors in the program.

Code

The code below shows how to obtain a monotone missing data pattern for the SLEDAI-2K total score.

```
proc mi data= sle_data seed=123456 nimpute=10 out= imp_int_sle;
    mcmc chain = multiple impute=monotone;
    var grp resp_1 resp_2 resp_3;
run;
```

The resulting output dataset imp_int_sle will have 10 copies of the original dataset where each copy will have a monotone missing pattern.
The code below shows how to use PROC MI by predictive mean matching method.

```
proc sort data= imp_int_sle; by _Imputation_;run;
proc mi data = imp_int_sle seed = 123456 nimpute = 1
    maximum=. . . 105 105 105
    minimum=. . . 0 0 0
    round = . . . 1 1 1
    out = sri_imp;
    by _Imputation_;
    class grp factors;
    Monotone regpmm (resp_1 = grp factors base);
    Monotone regpmm (resp_2 = grp factors base resp_1);
    Monotone regpmm (resp_3 = grp factors base resp_1 resp_2);
    var grp factors base resp_1 resp_2 resp_3;
run;
```

Through this program, we will generate the completed datasets. The imputed values should follow the objective evaluation results, so the limits of maximum value, minimum value and decimal places should be added in the code.

Meanwhile, several issues often arise during the imputation process, such as:

- Capping imputed values in limits: Although we have added upper and lower limits during the imputation process, to avoid values exceeding these limits, it is necessary to handle the results that exceed the upper and lower limits.

The imputation method for the PGA score is the same as that for the SLEDAI-2K total score, except that its upper limit is 3.

Step.2. - binary variable

No new A organ domain score or no more than 1 new B organ domain scores compared with baseline in BILAG-2004 is also a binary variable, so the imputation method is consistent with that for SRI-4 response. However, it requires additional consideration of the impact of baseline data on subsequent evaluations. The currently common approach is to grade the BILAG-2004 scores at baseline to obtain a quantifiable value.

Step.3.

We have now obtained the completed datasets for three indicators. After merging the imputed results of the three indicators, SRI-4 is derived according to preset rules—where a subject is classified as a responder only if all three individual indicators simultaneously yield positive assessments.

RESULTS

This article conducts a descriptive summary of the rate differences between groups calculated by two methods through 50 random simulations.

Based on a simple summary of the results (Table 1), it was found that the differences between the two methods are not substantial, the results obtained from Method 1 are closer to the adjusted CMH difference rate of the complete dataset(18.6%).

Methods	N	MIN	MAX	MEAN	STD	MEDIAN	Q1	Q3
METHODS 1	50	11.1	26.1	18.88	3.014	19.13	16.99	20.36
METHODS 2	50	8.4	25.9	19.00	3.367	18.82	17.75	21.11

Table 1. Summary of results for different methods

To further investigate the stability of the results, simulations were conducted with data missing rates of 8% and 12%. The findings revealed changes in the outcomes obtained by the two methods. When the missing rate was 8%, Method 1 produced more stable results that were closer to the difference rate of the complete dataset, as shown in Table 2. When the missing rate was 12%, Method 1 produced more stable results that were closer to the difference rate of the complete dataset, as shown in Table 3.

Methods	N	MIN	MAX	MEAN	STD	MEDIAN	Q1	Q3
METHODS 1	50	11.9	24.7	18.71	2.302	18.64	17.16	20.07
METHODS 2	50	12.3	25.6	18.78	2.752	19.03	17.17	20.22

Table 2. Summary of results for different methods when missing rates of 8%

Methods	N	MIN	MAX	MEAN	STD	MEDIAN	Q1	Q3
METHODS 1	50	11.2	25.8	18.70	2.820	18.86	17.05	20.03
METHODS 2	50	10.7	24.2	18.89	3.239	18.82	16.84	21.66

Table 3. Summary of results for different methods when missing rates of 12%

CONCLUSION

The estimatied results of both methods did not significantly deviate from the adjusted CMH difference rate of the complete dataset(18.6%). Method 1 exhibited smaller deviations from the true value under low missing rates (8%) and demonstrated more stability advantages.

Both methods provided generally reliable estimation of rate difference. Yet, Method 1 achieved higher precision and stability—closer to the true value—when data completeness was ensured (e.g., with low missing rates). In practical applications, Method 1 is recommended to minimize estimation bias.

It should be noted, however, that the instability of Method 2 may stem from the larger proportion of data imputation required when all corresponding subclass results were lost during missing data processing. In subsequent research, scenarios where all three indicators are completely missing or partially missing can be further explored through comparative studies. The imputation process is influenced by numerous factors, and the findings presented here should be taken as preliminary results.

REFERENCES

Michael O'Kelly, Bohdana Ratitch (2014), *Clinical Trials with Missing Data: A Guide for Practitioners*, John Wiley & Sons.

Fung Lam 1, Chi Chiu Mok (2023), Disease Activity Indices in Systemic Lupus Erythematosus: What Is New?, *JOURNAL OF CLINICAL RHEUMATOLOGY AND IMMUNOLOGY*, 23:60 – 67

Ratitch, B. and O'Kelly, M. (2013), "Combining Analysis Results from Multiply Imputed Categorical Data" in *Proceedings of PharmaSUG 2013 (Pharmaceutical Industry SAS Users Group)*, SP03, Nashville.

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

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *SAS® For Dummies®*
- *Flexible Imputation of Missing Data, Second Edition*

CONTACT INFORMATION

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

<Name: Binhan Chen>
<Enterprise: InnoCare Pharma>
<E-mail: Binhan.Chen@innocarepharma.com>