Paper AS01

Multiple Imputation in Composite Endpoints - It's Not That Hard!

Ashley (Huiyan) Mao, Roche, Mississauga, Canada

ABSTRACT

Missing data in clinical trials is a common issue that can potentially introduce bias and imprecisions in results interpretation. Various approaches exist to address missing data, among which Multiple Imputation (MI) has become the preferred method due to its ability to provide more reliable and valid statistical inferences under realistic assumptions of the missingness mechanism. MI takes into account the uncertainty associated with missing data by creating multiple filled-in datasets, thus yielding more precise estimates of standard errors and confidence intervals.

While MI is effective in missing data imputation for continuous or binary endpoints, its application in composite endpoints is not always straightforward. Careful consideration is required in model and variable selection, pooling strategy and result interpretation. In this presentation, we will illustrate in a case study how we apply MI techniques to handle the missing data in composite endpoint.

INTRODUCTION

Missing data is commonly observed in clinical trials, occurring when planned data is not available. The primary reasons for missing data include patients lost to follow-up, data collection error and missed scheduled visits. Missing data could impose greater impact in certain types of studies such as longitudinal studies due to repeated measurement across time. If not appropriately addressed, missing data can lead to biased conclusions and undermine the validity of study findings.

One simple approach to handle missing data is single imputation, where sample mean or median is used to impute the missingness. One caveat is that this method assumes the overall sample is representative of the patient with missing data, which may not always hold true. A similar but improved method is to perform single imputation based on covariates or regression model. For example, imputing missing values using treatment information or medical history helps improving the predictability of the imputed values. In the context of longitudinal dataset, Last Observation Carried Forward (LOCF) method is commonly used, where previous observations are leveraged to impute later visits, assuming the outcome is stable from the last visit.

While single imputation is simple and straightforward to implement, it does have significant limitations including strong assumptions and potential bias. A more sophisticated approach to manage missing data is Multiple Imputation (MI), which improves robustness and statistical inferences. This paper will present the general procedure for MI and a case study to explore the application of MI in composite endpoints.

MULTIPLE IMPUTATION - PROCEDURE

The overall procedure for MI involves three steps - imputation, analysis, and pooling. In the imputation step, the entire dataset will be generated multiple times (M), with missing data getting populated each time. The analysis model will be applied in each of the M datasets to obtain M estimates. In the final pooling step, M estimates will be pooled to generate a combined result.

Imputation step

There are several imputation methods available, and this paper will focus on a widely used technique known as Fully Conditional Specification (FCS). Unlike methods that heavily rely on a pre-specified joint distribution, FCS performs imputation by specifying conditional distributions. Let Y_j represents the observed data for the jth variable, where $j = 1,\ldots,p$ represents the order of the variables. Additionally let β_j denote the parameters of the distribution where Y is drawn. FCS iteratively estimates the parameter β, followed by imputing missing value of Y_j conditioned on β_j and other variables. Let t represents the iteration of the imputation process, where $t = 1,\ldots,M$. The algorithm is specified below:

1) Estimate parameters from posterior distribution
$$\beta_1^t \sim P(\beta_1 | Y_{1(obs)}, Y_2^{t-1}, \dots, Y_p^{t-1})$$
2) Fill in the missing value from observed data conditioned on the parameter β_1^t
$$Y_{1(*)}^t \sim P(Y_1 \mid \beta_1^t)$$
3) Variable Y_1 now consists of observed value $Y_{1(obs)}$ and imputed value $Y_{1(*)}^t$ at t-th iteration.
$$Y_1^t = (Y_{1(obs)}, Y_{1(*)}^t)$$
4) Move to Y_2 and repeat the same process to obtain imputed value at t-th iteration
5) Each of the variables j = 1,...,p is imputed in sequence at t-th iteration
6) Repeat the same process from step 1 in all M iterations

Depending on the property of each variable, different imputation models could be selected. For imputation of continuous variables, regression Predictive Mean Matching (PMM) is a popular method and is used in the case study discussed in this paper. PMM uses the parameters and other variables to obtain the predictive value for variable Y_j. For each Y_j that is missing, a pre-specified number of donors, whose predicted values are closest to those for missing Y_j, are selected. A random draw will be performed to fill in the missing data for Y_j with observed data copied from that donor. The process is repeated for each subject with missing Y_j. This method has been used widely as it does not require normality assumption thus it works with skewed data and it is more robust to model misspecification. For binary variables, logistic regression is used to impute the missing Y_j based on all the variables specified in the model.

The above algorithm will be run separately in each of the treatment arms to account for the difference in outcome attributable to the treatment. Depending on the level of missingness, the number of iterations M is usually between 20-100. The number of donors also needs to be handled with care. The number is typically chosen between 3, 5 or 10, which ensures enough variability while reducing the probability of mis-match.

Analysis step

At the end of the imputation step, M sets of completed datasets will be generated from each treatment arm. In this step, complete datasets will be combined from both treatment arms to fit the analysis model from each of M datasets. Based on the pre-specified analytical method, the outcome of interest is estimated from M datasets, thus obtaining M different estimates θ, and their associated variance W.

Pooling step

Under Rubin's rule, the pooled point estimate of outcome of interest is calculated as the average of M estimates following the calculation below:
$$\hat{\theta} = \frac{1}{M}\sum_{t=1}^{M} \hat{\theta}_t$$

The standard error of the pooled estimate is composed of within imputation variance and between imputation variance, and is calculated as follows:
$$W_{total}(\theta) = \sum W_{within}(\theta) + (1 + \tfrac{1}{M}) W_{between}(\theta),$$
where within variance is $W_{within} = \frac{1}{M}\sum_{t=1}^{M} W_t$ and between imputation variance is $W_{between} = \frac{\sum_{t=1}^{M}(\theta_t - \hat{\theta})^2}{M-1}$.
The transformed statistics $\frac{\hat{\theta}}{\sqrt{W_{total}(\theta)}}$ follow a t-distribution where degree of freedom is derived using the formula below:
$$df = (M-1)[1 + \frac{W_{within}(\theta)}{(1+\frac{1}{M}) W_{between\ (\theta)}}]^2$$

APPLICATION OF MULTIPLE IMPUTATION IN COMPOSITE ENDPOINT

What is a composite endpoint?

MI is commonly applied to continuous or binary variables, but its application to composite endpoint is less straightforward. A composite endpoint is a single measure of effect that combines outcomes from multiple individual endpoints. This approach

provides a more comprehensive evaluation of disease activity, especially in disease areas with complexity and wide range of clinically relevant outcomes.
The application of MI to composite endpoints could be challenging for several reasons. The components might have different missing data patterns or are composed of different types of data. In addition, an important consideration is whether to perform imputation at composite or component level. Imputing at component level requires specifying the correlation between components, while at composite level data for components might not be considered. In the next paragraphs, the paper will talk about the implementation of MI in composite endpoint in a case study.

End-to-end imputation and analysis process

The use of the Systemic Lupus Erythematosus Responder Index (SRI) in lupus research is an example of composite endpoint. Patients are classified as either responders or non-responders, based on whether or not he/she has met the individual components of SRI. For illustration purposes, we use the composite endpoint SRI as an example. SRI is comprised of the following components:

- Measure 1 (continuous)
- Measure 2 (continuous)
- Measure 3 (binary)

Each of the components is measured in a longitudinal manner. The following An example dataset *dat* looks like the following:

SUBJID	TRTARM	REGION	AGE	AVISIT	PARAM	AVAL
1010-1001	A	1	39	Baseline	Measure 1	18
1010-1001	A	1	39	Week 4	Measure 1	18
1010-1001	A	1	39	Week 8	Measure 1	16
1010-1001	A	1	39	Week 12	Measure 1	8
1010-1001	A	1	39	Baseline	Measure 2	3
1010-1001	A	1	39	Week 4	Measure 2	3
1010-1001	A	1	39	Week 8	Measure 2	2
1010-1001	A	1	39	Week 12	Measure 2	0
1010-1001	A	1	39	Baseline	Measure 3	N
1010-1001	A	1	39	Week 4	Measure 3	N
1010-1001	A	1	39	Week 8	Measure 3	Y
1010-1001	A	1	39	Week 12	Measure 3	Y
1010-1002	B	2	42	Baseline	Measure 1	10
1010-1002	B	2	42	Week 4	Measure 1	12
1010-1002	B	2	42	Week 8	Measure 1	11
1010-1002	B	2	42	Week 12	Measure 1	12
1010-1002	B	2	42	Baseline	Measure 2	2
1010-1002	B	2	42	Week 4	Measure 2	3
1010-1002	B	2	42	Week 8	Measure 2	3
1010-1002	B	2	42	Week 12	Measure 2	1
1010-1002	B	2	42	Baseline	Measure 3	N
1010-1002	B	2	42	Week 4	Measure 3	Y
1010-1002	B	2	42	Week 8	Measure 3	Y
1010-1002	B	2	42	Week 12	Measure 3	Y

…….

Before implementing the imputation step, the dataset is pre-processed to wide format and examined for missing data patterns. The data preprocessing can be implemented using the following R code:

```
dat_w <- dat %>%
  spread(AVISIT, AVAL)
```

Partial data for the wide dataset *dat_w* now looks like the following:

SUBJID	TRTARM	REGION	AGE	PARAM	Baseline	Week 4	Week 8	Week 12
1010-1001	A	1	39	Measure 1	18	18	16	8
1010-1001	A	1	39	Measure 2	3	3	2	0
1010-1001	A	1	39	Measure 3	N	N	Y	Y
1010-1002	B	2	42	Measure 1	10	12	11	12
1010-1002	B	2	42	Measure 2	2	3	3	1
1010-1002	B	2	42	Measure 3	N	Y	Y	Y
1010-1003	A	2	35	Measure 1	14	18	14	10
1010-1003	A	2	35	Measure 2	3	2	<NA>	<NA>
1010-1003	A	2	35	Measure 3	N	Y	<NA>	<NA>
1010-1004	B	2	29	Measure 1	12	<NA>	<NA>	<NA>
1010-1004	B	2	29	Measure 2	2	2	<NA>	<NA>
1010-1004	B	2	29	Measure 3	N	N	N	<NA>

……

Based on the missing data pattern, the data is found to be missing monotonically and is assumed to be missing at random (MAR), thus it's appropriate to use FCS method for imputation. The paragraphs below illustrate the 3 steps involved from the imputation step to final estimation of the endpoint.

Imputation step

Imputation is performed at each component level to preserve the observed data. Based on the missing patterns and the nature of the variables, PMM is selected for Measure 1 and Measure 2, and logistic regression is selected for Measure 3. Selection of covariates in the imputation model takes into consideration of clinical relevance and correlation of the outcome variables. In our case study, each of the components were measured every 4 weeks. It was thought that age and region are the key demographics in the disease outcome. Measurements from previous visits are predictive of later visits. Measure 1 and Measure 2 are found to be correlated. Therefore, the imputation model is set up as follows:

$$Measure\ 1_j = \beta_0 + \sum_{r=0}^{j-1} \beta_{r+1} \times Measure\ 1_r + \sum_{r=0}^{j-1} \beta_{r+j+1} \times Measure\ 2_r + \beta_{2j+1} \times age + \beta_{2j+2} \times region$$

$$Measure\ 2_j = \beta_0 + \sum_{r=0}^{j-1} \beta_{r+1} \times Measure\ 2_r + \sum_{r=0}^{j-1} \beta_{r+j+1} \times Measure\ 1_r + \beta_{2j+1} \times age + \beta_{2j+2} \times region$$

$$Measure\ 3_j = \beta_0 + \sum_{r=0}^{j-1} \beta_{r+1} \times Measure\ 3_r + \beta_{j+1} \times age + \beta_{j+2} \times region$$

where j = 1,…,3 such that measurements are performed at Week j x 4.

Separate imputation models are run within arm A and B. The imputation is repeated 100 times. Each time the missing value is replaced with an observed value from a donor with proximity, creating 100 complete datasets. The code can be found below:

```
### For Measure 1 and Measure 2 (continuous)
imputed_armA <- mice(data = dat_w, method = "pmm", m = 100, predictorMatrix = pm, donors = 5)
imputed_armB <- mice(data = dat_w, method = "pmm", m = 100, predictorMatrix = pm, donors = 5)

### For Measure 3 (binary)
imputed_armA <- mice(data = dat_w, method = "logreg", m = 100, predictorMatrix = pm)
imputed_armB <- mice(data = dat_w, method = "logreg", m = 100, predictorMatrix = pm)
```

The predictor matrix *pm* is a matrix consisting of 0 and 1, where 1 indicates the variable is used in the imputation model and 0 indicates otherwise. By employing the predictor matrix, the imputation model is specified.

Analysis & pooling step

In the 100 iterations, all 3 components at all visits are filled. Since the composite endpoint - SRI is the endpoint of interest, SRI response is derived for each iteration M. Therefore, a binary SRI response is created for each of the 100 imputations and for each subject. Odds ratio is calculated for each iteration, obtaining 100 estimates and their associated variance.

Since Rubin's rule relies on normal assumption, the odds ratio values are log-transformed and averaged to obtain a pooled statistic *log_odds_ratios_avg*. Standard error is also normalized following the formula $\frac{(\log(CI_{upper})-\log(CI_{lower}))}{2 \times 1.96}$. The transformed statistics follow a t-distribution with degree of freedom *dt_t* based on *M* and between (*B_v*) and within (*W_v*) imputation variance. Total variance *T_v* is calculated based on *B_v* and *W_v.* In the end t statistic is calculated using the formula $\frac{log\_odds\_ratios\_avg}{\sqrt{T\_v}}$ for statistical inferences. The R code looks like below:

```
# Loop over each imputed dataset
for (i in 1:100) {

  # Subset imputation iterations
  imputed_dat <- complete(imputed_dataset, action = i)

  # Fit logistic regression model
  model <- glm(SRI4 ~ TRTARM, data = imputed_dat, family = binomial)

  # Extract log odds ratio
  log_odds_ratios[i] <- coef(model)["TRTARM"]

  # Extract standard error for the log odds ratio
  std_error <- summary(model)$coefficients["TRTARM", "Std. Error"]
  ci_l <- confint(model)["TRTARM", ][1]
  ci_u <- confint(model)["TRTARM", ][2]
  std_error_norm <- (log(ci_u) - log(ci_l))/(2*1.96)

  # Calculate variance
  variances[i] <- std_error_norm^2
}
# Pool log odds ratio
log_odds_ratios_avg <- mean(log_odds_ratios)

# Obtain between-imputation variance and within-imputation variance
B_v <- sum((log_odds_ratios - log_odds_ratios_avg)^2) / (100 - 1)
W_v <- mean(variances)

# Calculate total variance
T_v <- W_v + B_v*(1 + 1/100)

# Calculate degrees of freedom for t distribution
df_t <- (100 - 1) * (1 + W_v/((1+1/100) * B_v))^2
```

CONCLUSION

MI represents a robust statistical technique for missing data handling. By leveraging MI, uncertainty and randomness have been accounted for, which mitigates the bias associated with missing data. However, implementation of MI could be challenging, especially when dealing with composite endpoints. It also requires extra consideration in the interpretation of results. Therefore, it is recommended to report the observed data alongside the imputed outcomes when presenting the results. Given the growing preference for MI over single imputation such as LOCF, early interactions with the Health Authorities are encouraged. It is also essential to pre-specify the imputation models and justify the choice of parameters in the Statistical Analysis Plan (SAP).

This paper demonstrates the approach to handle missing data in composite endpoints with binary outcomes. The methods can certainly be extrapolated for future analysis of composite endpoints with continuous outcomes or with various types of components.

REFERENCES

Kang, H. (2013a). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402. https://doi.org/10.4097/kjae.2013.64.5.402

Buuren, S. van. (2021). *Flexible imputation of missing data*. Chapman & Hall/CRC.

Liu, Y., & De, A. (2015). Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. *International Journal of Statistics in Medical Research*, *4*(3), 287–295. https://doi.org/10.6000/1929-6029.2015.04.03.7

Pham, T. M., White, I. R., Kahan, B. C., Morris, T. P., Stanworth, S. J., & Forbes, G. (2021). A comparison of methods for analyzing a binary composite endpoint with partially observed components in randomized controlled trials. *Statistics in Medicine*, *40*(29), 6634–6650. https://doi.org/10.1002/sim.9203

O'Kelly, M., & Ratitch, B. (2014). *Clinical Trials with Missing Data*. https://doi.org/10.1002/9781118762516

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W. J., Siegel, J. P., & Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, *367*(14), 1355–1360. https://doi.org/10.1056/nejmsr1203730

ACKNOWLEDGMENTS

I would like to express my gratitude to the working group at Roche: Armando Turchetta (former employee), Ben Lanza, Claire Hughes, Eriola Berisha, Hao Wang, Imran Hassan, Jennifer Pulley, Lynna Zhao, Pascal Guibord. Their expertise, dedication and enthusiasm have been instrumental in driving this research forward. It's my pleasure to work alongside such a great team.

CONTACT INFORMATION

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

Author Name: Ashley (Huiyan) Mao
Company: Roche Canada
Email: ashley.mao.am1@roche.com