Using Planned Missing Values in Longitudinal Trials to Relieve Patient Burden and Reduce Costs

Robert D. Small, PhD[1], Ayca Ozol-Godfrey, PhD[1],
Dominika M. Wisniewska, MS[2], Christele Augard, MS[2]
[1] Sanofi Pasteur, Discovery Drive, Swiftwater, PA, 18370
[2] Sanofi Pasteur, 1541 avenue Marcel Merieux – 69280 MARCY L'ETOILE - France

Abstract

Some longitudinal trials require subjects to submit to frequent blood draws on visits over time. Often the primary endpoint does not require the observations from every visit. This is true, for example, in vaccine immunogenicity trials and in diabetes trials. Taking samples at every visit can be burdensome to both the subject and the sponsor. Subjects often do not like many blood draws. The cost of assays of every sample can be high. These facts contribute to increased cost and increased subject drop out. In this paper we investigate the idea of bleeding random subsamples at each visit but using the (frequent) high correlation within subjects between visits to build imputation models to implement an MI approach to analyzing the data. We use the observations present as well as other pertinent continuous and categorical variables to build the models. We do the estimation of the imputation models using a method of Raghunathan, Lepkowski et.al. which is very general and can handle many types variables. We give examples using data from some recent vaccine trials. We show how various patterns can reduce cost and possibly drop outs.

Key Words: missing data, trial design, sample size

1. Introduction

Frequent blood draws in some clinical trials can be burdensome to the subjects and to the sponsors. They are associated with an increased cost based on the blood sample handling, transferring, and testing. For example, assaying a blood sample in a vaccine (1) trial can cost from $50 to $200 depending on the test. In addition, too many blood draw visits can simply cause subjects to drop out of trials, especially in the pediatric and elderly populations. If only a random subset of the subjects could be bled at each physical site visit, this could possibly reduce cost and the drop-out rate.

With this approach, the statistician can still use the high/frequent correlations between visits to build imputation models to implement a multiple imputation approach to analyze the data. In this paper, we use the sequential regression multivariate imputation (SRMI) of Raghunathan, Lepkowski, et.al. (2) to perform the imputation. In addition standard multiple imputation (MI) is used as a comparison.

With this approach, for the final analysis, the imputed results would be considered as the primary results. This concept in a way prevents missing data due to drop-outs and still gives a way to analyze the data in a statistically meaningful way.

In this paper, in section 2 the study design is presented and the two methods: sequential regression multivariate imputation (SRMI) and multiple imputation (MI) are described. In Section 3 the details of the two examples are presented along with the results. Section 4 is a short discussion.

2. Study Design and Methods

2.1 Study Design

This was a Phase II, randomized, mono-center vaccine trial in 2 to 11 year old children in Peru. Each subject was to receive 3 vaccinations of either the active or the control group at Days 0, 180, and 365. The planned sample size was 200 and 100 for the active group and the control group, respectively. The aim of the trial was to describe the serological immune response to the disease virus before and after each vaccination, 28 days post-dose 3 being our primary time point. Additionally, presence of viremia, the level of disease vaccine virus, was to be assessed. Since viremia was expected to occur at low levels, it was chosen to be assessed twice, 8 and 15 days after the first and the second vaccination. Since in this trial, viremia presence was very low after the second vaccination for the active group and 0% for the control group, we only considered the presence of viremia at either Day 8 or Day 15 after the first vaccination for the active group. In order to limit the blood volume taken from the children, the viremia assessment was made only in a random subgroup of subjects (active group=97, control group=30 subjects).

Table 1 shows the vaccination schedule, the viremia and the serology blood draw visits.

Table 1: Trial flow chart[a]

Visit	V01	V02	V03	V04	V06	V07	V08	V09	V11	V12
Days	D0	D8	D15	D28	D180	D180 +8	D180 +15	D180 +28	D365	D365 +28
Viremia		BL2	BL3			BL6	BL7			
Serology	BL1			BL4	BL5			BL8	BL9	BL10
Vaccine	Vax 1				Vax 2				Vax 3	

In this vaccine trial, a laboratory tested the serology blood samples taken from each subject at pre-vaccination bleeds 1, 5, and 9, and post-vaccination bleeds 4, 8, and 10 using a microneutralization assay. The antibody response was reported in titers. In this paper, we are only interested in a single serotype, but the correlations between serotypes can also be investigated to come up with an imputation approach.

[a] V=Visit, D= Day, BL=Bleed, Vax=Vaccination

2.2 Methods

In this paper, we use the sequential regression multivariate imputation (SRMI) method of Raghunathan, Lepkowski, et.al.(2) to impute missing values. This is a very general imputation method which can handle many different variable types, in our case categorical and continuous. In addition, it can handle restrictions like the fact that when the log10 (titer) is imputed, it has to be above 0. A SAS based application is readily available by the authors on the web from the University of Michigan. In addition, we use standard multiple imputation to compare the results. Once the imputed datasets (m=50 datasets) are obtained by each technique, each dataset is analyzed separately, and the estimates and the standard deviations are combined (Rubin, 2002, [3]) to obtain the final results.

The logarithm (in base 10) of the titer is the variable included in the models and the imputations, and when the analysis is complete, the anti-logs were taken and the geometric mean titers (GMTs) were reported. In this paper, we reported the individual group GMTs and the GMT ratios of active versus control group comparisons.

2.2.1 Sequential Regression Multivariate Imputation (SRMI)

Sequential regression multivariate imputation is an imputation procedure that can handle complex data structures by combining regression and Bayesian ideas. Basically, it obtains imputations by fitting a sequence of regression models and drawing values from the corresponding predictive distributions. It has an ability to accommodate complex data structures using normal linear regression, logistic regression, generalized logistic regression, Poisson regression, and two stage models based on zero-non zero status.

Suppose U is the set of the completely observed variables that are a mixture of continuous, binary, count or mixed variables in your imputation model. $Y=Y_1 ,\ldots,Y_p$ denotes the p variables with missing values, ordered assuming Y_1 has the least number of missing values and Y_p has the most number of missing values. The first iteration starts by regressing Y_1 on U, imputing the missing values given the regression model and obtaining a posterior predictive distribution of Y_1 given U. Then Y_2 is regressed on (U, Y_1), which gives the posterior predictive distribution of Y_2 given U and Y_1. The cycle continues through these series of regression models until the last Y variable is imputed. Once Y_p is imputed the first iteration is complete. The imputations continue as $Y_1,\ldots,Y_p$ are alternately regressed on all the other variables, i.e. regress Y_1 on (U, $Y_2 ,\ldots Y_p$); regress Y_2 on (U, Y_1 , $Y_3,\ldots,Y_p$); etc., and impute draws from posterior prediction distributions thus updating the imputed values. The imputations continue in a cyclic manner, each time overwriting the previously drawn values, until convergence is reached. SRMI is a very flexible method in the sense that it can accommodate complex data features using different regression methods. Moreover, the data does not need to have an explicit joint multivariate distribution of all the variables. This algorithm can be computationally intense when large and complex data sets are of interest. However, the algorithm can be modified to apply a variable selection method for each regression in each round.

The authors have developed a SAS based application called Imputation and Variance Estimation Software (IVEware) to implement this approach, which is available from a website (www.isr.umicj.edu/src/smp/ive).

2.2.2 Multiple Imputation (MI)

MI is a technique where each missing value is replaced with a set of values that represents the uncertainty about the right value to impute. The basic idea of MI in a regression setting is to use a complete dataset as linear regression predictors to estimate the distributions of the regression coefficients and then use these regression coefficients to predict and impute the missing values.

We have used SAS (4) proc MI and proc MIANALYZE to do our imputations. Because proc MI requires a monotone missing pattern (i.e. when Y_j is missing for an individual, then it is assumed that all subsequent variables Y_k, k>j are missing for that individual) and in vaccine trials most often there is a small amount of missing data at the pre-blood draws, we first imputed enough number of missing values using a Markov chain Monte Carlo (MCMC) method to make the imputed datasets have a monotone missing pattern.

3. Examples and Results

3.1 Example 1

In this first example, our primary point of interest was the presence of viremia in the active group, a categorical variable. Table 2 shows the different imputation models and methods that were used. In the observed complete data, the presence of viremia was 44.3% (43/97 subjects). We used 2 different regression models:

1. Presence of viremia (Yes/No) = Bleed 1+Bleed 4+Bleed 5+Bleed 8
to investigate the relationship between viremia and the serological immune response

2. Presence of viremia (Yes/No) = Age+BMI+YF+Bleed 1+Bleed 4+Bleed 5+ Bleed 8
to investigate the relationship between viremia and the serological immune response along with some covariates.

Table 2: Models used in the viremia example

N	Model[b]	Model #	Method	Minrsq
97	Observed viremia rate (43 out of 97 subjects: 44.3%)	0	Observed	N/A
199	Detectable viremia (Yes/no)= BL1+BL4+BL5+BL8	1	IVE	None
199	Detectable viremia (yes/no)= Age+BMI+Sex+YF+BL1+BL4+BL5+BL8	2	IVE	None
199	Detectable viremia (Yes/no)= BL1+BL4+BL5+BL8	1	IVE	0.01
199	Detectable viremia (yes/no)= Age+BMI+Sex+YF+BL1+BL4+BL5+BL8	2	IVE	0.01
199	Detectable viremia (Yes/no)= BL1+BL4+BL5+BL8	1	IVE	0.25
199	Detectable viremia (yes/no)= Age+BMI+Sex+YF+BL1+BL4+BL5+BL8	2	IVE	0.25
199	Detectable viremia (Yes/no)= BL1+BL4+BL5+BL8	1	MI	N/A

[b] BMI=Body mass index, YF= Yellow fever serological status, BL=Bleed
BL1, BL5 = Pre-dose 1 and pre-dose 2 serology bleeds, respectively
BL4, BL8= 28 days post-dose 1 and post-dose 2 serology bleeds, respectively

199	Detectable viremia (yes/no)= Age+BMI+Sex+YF+BL1+BL4+BL5+BL8	2	MI	N/A

For both models, in addition a stepwise selection procedure was used to select the best predictors with the IVE software by setting a minimum marginal R-squared (min R-squared=0.01 and 0.25). Once the imputed datasets were obtained, the viremia percentage was obtained from a single binomial proportion.

Figure 1 shows the viremia percentages, along with the 95% CIs based on the different approaches. The 95% CI for the Observed viremia percentage based on the subset of 97 subjects is calculated using the exact binomial method (5). In addition, the numbers seen in the figure are the ratio of the CI widths with respect to the width of the Observed CI. As seen from the figure, the ratios of the CI widths are within decimals of each other for most cases. IVE-model 2 with the demographic covariates and no selection mechanism has the point estimate furthest from the observed, followed by the MI-model 2 estimate. As seen, it may make a difference what imputation model is used.

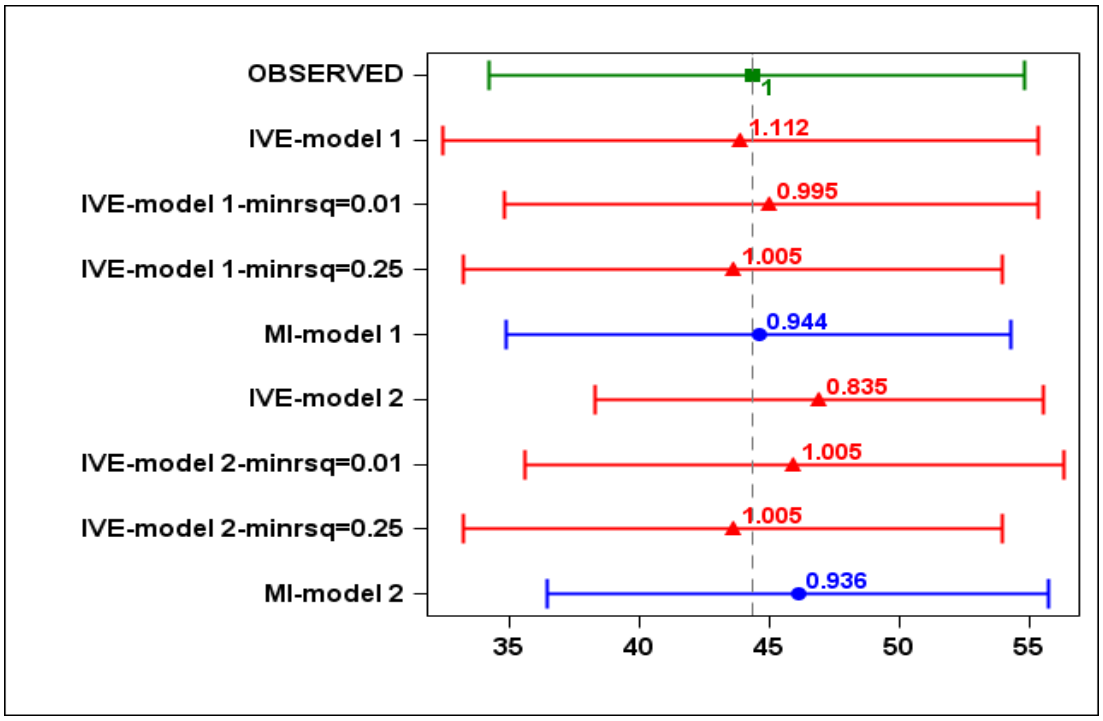

Figure 1: Active group: Viremia %s and 95% CIs (with ratio of CI widths with respect to "Observed" CI)

In addition, we looked at the post-dose individual GMTs of the active group. Figure 2 shows the 28 days post-dose 2 GMTs, along with the 95% CIs based on the different approaches similar to Figure 1. The Observed GMT in this figure is based on 188 subjects data as 11 subjects had a missing post-dose 2 visit out of the 199 subjects. As seen from the figure, for all the different models, the GMT point estimates were close to the observed GMT, within a -1.2 to 5.5% range. The CI widths were all within a -2.3% of the observed CI's widths. In addition, the IVE model with either no stepwise selection or min R-squared =0.1 and MI perform very similarly as well.

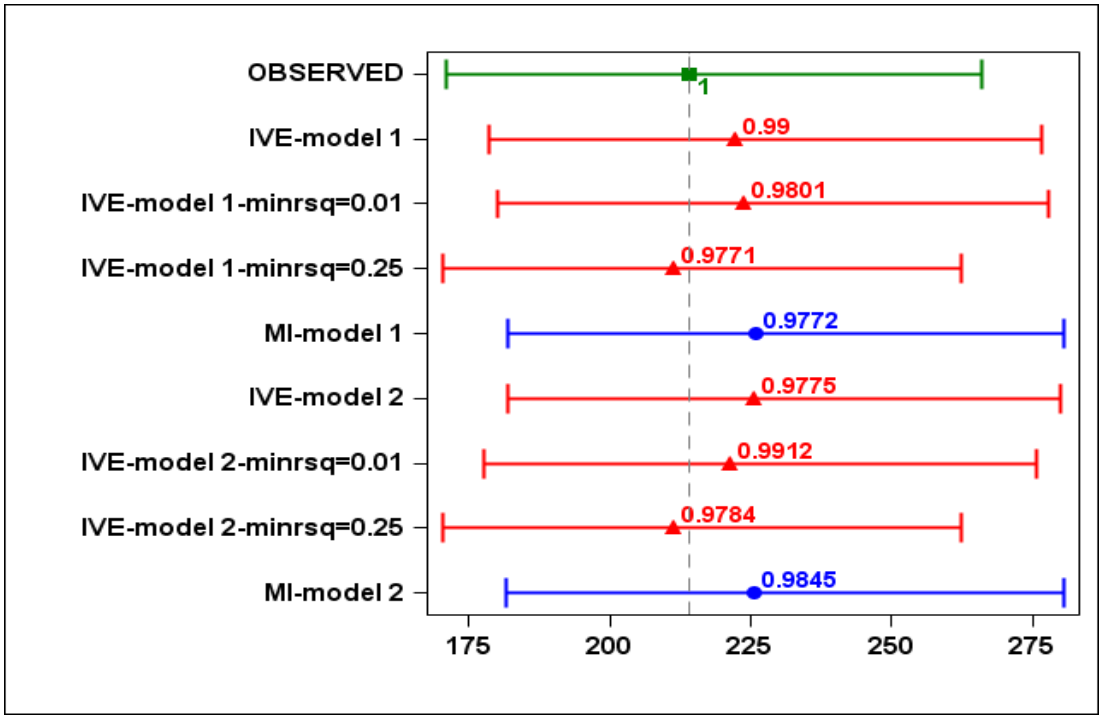

Figure 2: Active group: Post-dose 2 GMTs and 95% CIs (with ratio of CI widths with respect to "Observed" CI)

3.2 Example 2

In this second example, we apply SRMI and MI to the same vaccine trial in example 1 to compare the investigational vaccine to the control using the 28 days post-dose 3 GMT ratios. The post-dose 3 GMT ratio being our primary point of interest, gives the clinical team an opportunity to perhaps not collect blood samples from all the subjects at all their other previous time points.

In this example, the imputation model was Bleed 10 = Bleed 1+Bleed 4+Bleed 5+Bleed 8+Bleed 9 to investigate the relationship between the 28 days post-dose 3 and the other serology time points. We used the t-test based on the unequal variance assumption once the imputed datasets were created to obtain the estimates and the standard errors to merge.

Table 3 shows the variances and the Pearson correlation coefficient between the six serology visits' titers.

Table 3: Variances and Pearson correlation coefficients between serology visit[c] titers

Vaccine		Pre-dose 1	Post-dose 1	Pre-dose 2	Post-dose 2	Pre-dose 3	Post-dose 3
Active	Pre-dose 1	0.916	0.901	0.909	0.744	0.715	0.596
	Post-dose 1	0.901	1.175	0.939	0.792	0.741	0.566

[c] Pre-dose 1=BL1, 28 days post-dose 1=BL4, pre-dose 2= BL5, 28 days post-dose 2=BL8, pre-dose 3=BL9, 28 days post-dose 3=BL10

	Pre-dose 2	0.909	0.939	1.067	0.793	0.736	0.564
	Post-dose 2	0.744	0.792	0.793	0.419	0.735	0.660
	Pre-dose 3	0.715	0.741	0.736	0.735	0.962	0.784
	Post-dose 3	0.596	0.566	0.564	0.660	0.784	0.441
Control	Pre-dose 1	0.922	0.994	0.956	0.959	0.948	0.848
	Post-dose 1	0.994	0.922	0.959	0.962	0.948	0.846
	Pre-dose 2	0.956	0.959	0.970	0.997	0.978	0.874
	Post-dose 2	0.959	0.962	0.997	0.986	0.976	0.873
	Pre-dose 3	0.948	0.948	0.978	0.976	1.119	0.909
	Post-dose 3	0.848	0.846	0.874	0.873	0.909	1.184

As seen from Table 3, the correlations in the control group are all above 0.90 between the first five visits and still very high (>0.84) between the last visit and the previous visits. Because of the high correlations between visits, the loss of an observation at a single visit would result in only a small loss of information. For the active group, the correlations between the initial visits are very high and you can see a slight decrease after dose 2. However, even correlations around 0.5-0.6 at post-dose 3 with the previous visits are good enough if used properly.

Figure 3 shows the boxplots for both treatment groups at each serology time point. As expected for the active group, the individual GMTs increase after the first vaccination, decrease before the second vaccination, increase further after the second vaccination, and peek 30 days after the third vaccination. They are almost constant for the control group as expected. The slight increase at the third dose could be linked to disease exposure.

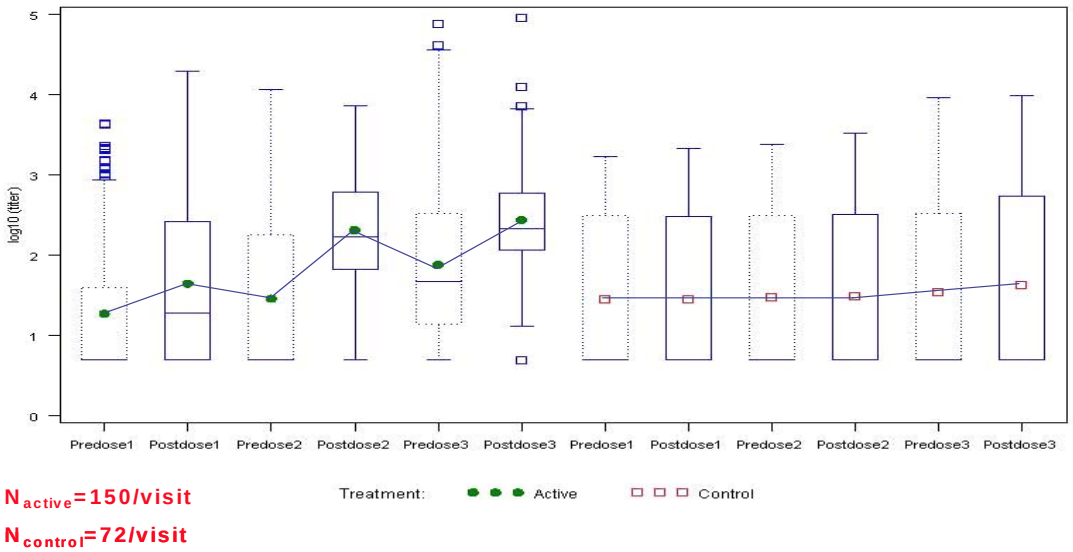

Figure 3: Serology time point boxplots for each treatment group

We took the complete dataset of 222 subjects with six serology visit titers and removed some of the subject visits' data using 3 different patterns. In the first two patterns, we removed titers in a systematic way. In pattern 1, one visit out of six for each subject was removed. In the second pattern, two visits were removed from each subject. For the third pattern, more observations were removed from the initial three visits (30% of titers/time point), slight less from visits 4 and 5 (25% of titers/time point), and the least (10% of titers/time point) was removed from the last visit, the 28 days post-dose 3 visit based on the correlations observed between visits as seen in Table 3. Figure 4 shows the display of the three patterns.

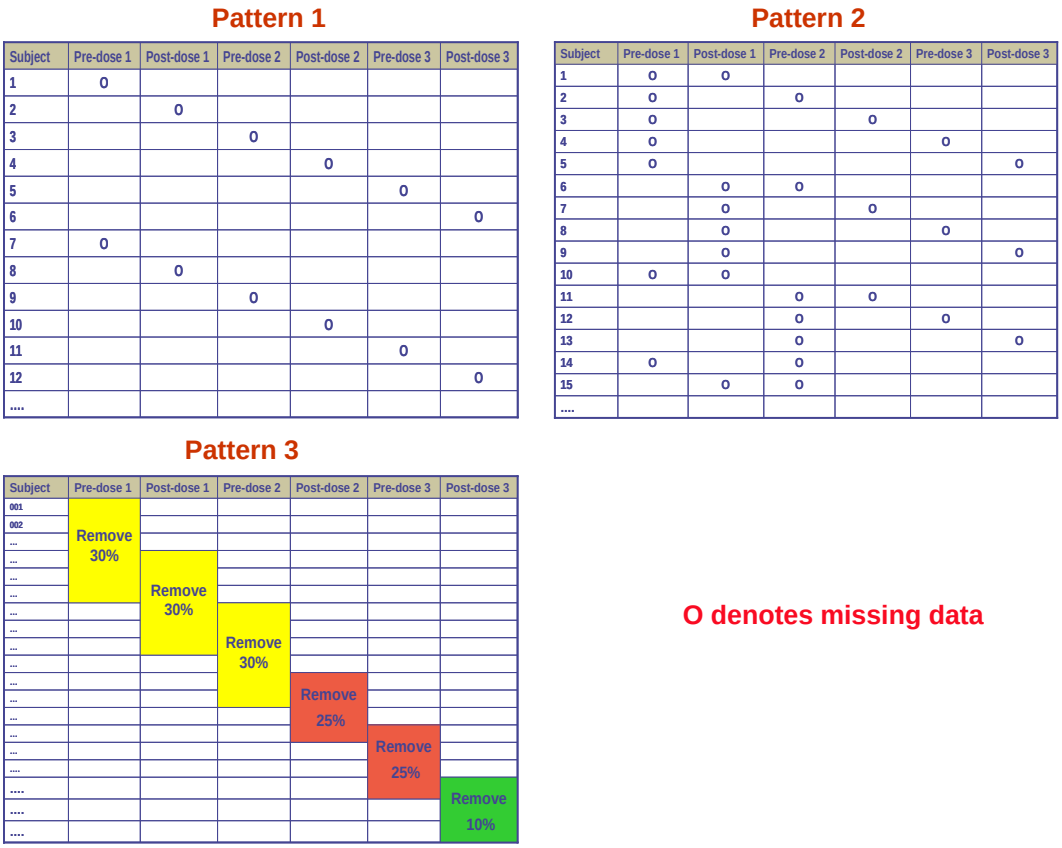

Figure 4: Missing data patterns

Overall, 16.7%, 33.3%, and 24.9% of the total 1332 observations were removed for each pattern, respectively. More important than the amount of observations removed is the amount of information removed. You will see indications of how much information was lost when the confidence intervals are compared below.

Figure 5 shows the 28 days post-dose 2 and post-dose 3 GMT ratios, along with their 95% CIs. The numbers seen in the figure are the ratio of the CI widths with respect to the width of the observed CI ("Observed" corresponds to the complete data before the titers were removed).

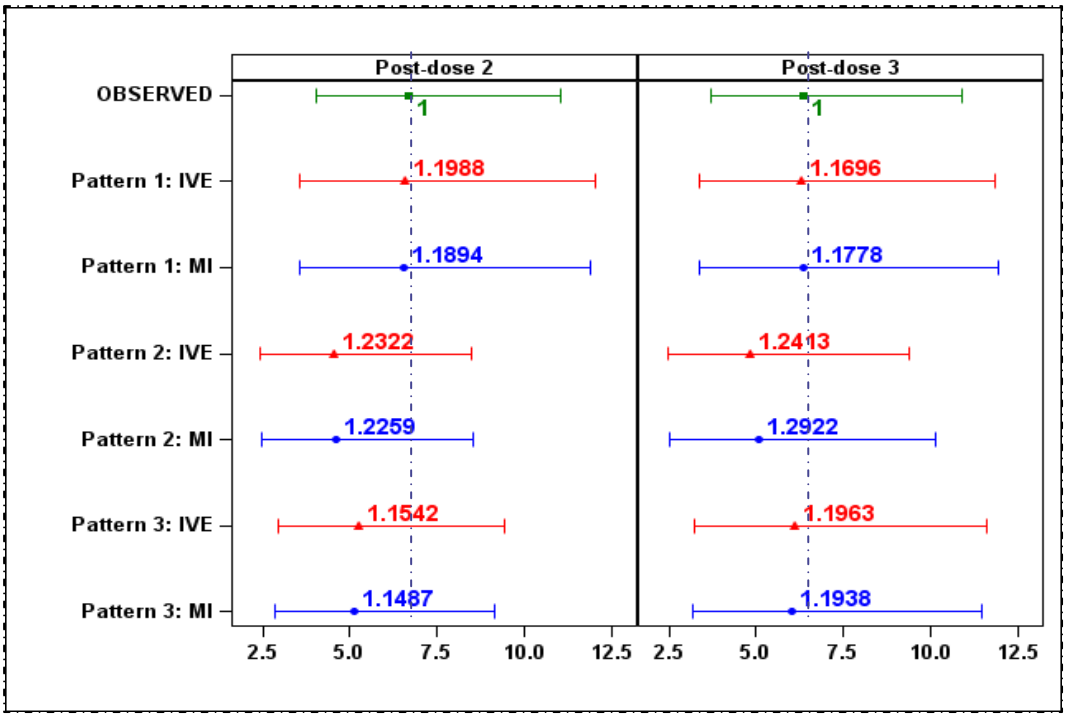

Figure 5: 28 days post-dose 2 and 3 GMT ratios and 95% CIs (with ratio of CI widths with respect to "Observed" CI)

As seen from Figure 5, for post dose 3, the CI widths are within 17% to 24% of the Observed CIs using SRMI. Pattern 2 has the widest CIs and pattern 1 has the narrowest CIs. Similar patterns were observed using MI. This may be due to the fact that for SRMI an imputation model with no stepwise selection approach was chosen. For post dose 2, pattern 3 has the narrowest CI widths followed by pattern 1 and pattern 2, ranging from 15% to 23% wider than observed for any method.

The point is, depending on the pattern one chooses to randomly remove the data, the optimum confidence intervals could be obtained. This idea could be used to design a clinical trial during the planning stages with the right assumptions.

4. Summary

In clinical trials that require frequent blood draws, such as vaccine and diabetes, it is possible to reduce the burden by decreasing the number of bleed visits based on the visit correlations during the planning stage. There should be an optimal way to obtain the best estimates based on a possible optimization algorithm. This way, we can still obtain estimates that are close to the actual estimate and CIs that are reasonable. As seen different patterns will produce different results based on the amount of information lost. When trying to decide how to do the removal process, the clinical trial team needs to carefully consider the assumptions. IVEware and MI are statistical tools to help us achieve this goal.

References

1. Nauta Jozef, Statistics in Vaccine Clinical Trials, Springer 2010
2. Raghunathan, T.E., et al. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models", Statistics Canada (2001), 27: 85-95

3. Little, Roderick J.A. and Rubin, Donald B., Statistical Analysis with Missing Data, Wiley 2002
4. SAS Institute Inc. 2008. SAS/STAT 9.2 User's Guide, Cary, NC, USA: SAS Institute Inc.
5. Newcombe R.G., "Two-sided confidence intervals for the single proportion: comparison of seven methods", Statistics in Medicine (1998) 17: 857-872