PhUSE 2012

Paper CS04

Macro to replace missing values

Kamrunnisa & Mokhfur Alam Chowdhury
Shafi Consultancy Bangladesh, Sylhet, Bangladesh

ABSTRACT

Missing data is common in most trials, whether it is random missing values in the raw data, or values which are excluded due to the use of rescue medication or for other causes. One thing for sure is that at least one method is usually specified in the analysis plan on what should be done with the missing values. Last Observation Carried Forwards (LOCF), Last Observation Carried Backwards (LOCB), Linear Interpolation and using summary statistic such as mean, median, minimum and maximum are the most widely used methods for replacing missing values. This paper will present a macro that can be used to replace missing values. This will help to reduce programming and validation time, as well as ensuring consistency both within and across studies. The macro can be used with different dataset structures after minor modifications, thus making it versatile and ensuring continuous future benefit.

INTRODUCTION

To capture all data from all patients is the objective in clinical trials. However, this does not happen often, resulting in missing values appearing in the data. Of course values are also excluded from analysis because rescue medication was used or other events had taken place. It should be noted that just ignoring these missing and excluded data is not an acceptable option when planning, conducting or interpreting the analysis of a confirmatory clinical trial. Fortunately, when there is missing data, some commonly used methods are available to replace missing values. This paper will show how programmers can save time, improve efficiency and consistency by the use of macro to replace missing values. The main purpose of the macro is to replace the missing values using one of four methods specified by the user in the macro call. The missing values can be imputed using last observation carried forward/backward, linear interpolation or by using a summary statistic, such as mean, median, mode, minimum or maximum values.

MACRO OVERVIEW

The macro 'IMPUTATION' starts by performing checks on the macro parameters used in the macro call to ensure they are consistent and valid. These are then used to determine what dataset is used, which variable contains the missing value, which method is used to replace the missing values and where the final data should be stored.

Validation check of parameters:

- **Input dataset validation check**: If the specified input dataset does not exist, a message is sent to the log and the macro will stop.
- **Validation for analyzed variables:** Check the analyzed variables that exist in the specified dataset or not.
- **Check variables those are being used as summary value:** Which summary values (i.e. MEAN, MEDIAN, MIN, MAX etc) will be used in place of missing values, before getting this value macro will check that user using the valid summary values by required parameter or not. Without getting right value macro will be stopped showing message in log.

Imputed methods validation: One of the following method must be specified with required parameters:

LOCF/LOCB : LOCF or LOCB - the last measured observation before the missing value is forwarded. This method works best if the observations are expected to remain at same level or if there are only a few missing values.

Baseline Observation Carried Forward is another single imputation approach that is sometimes used. To replace missing by baseline value, an additional parameter must also be specified in LOCF method to get baseline and replace missing by pre treatment value.

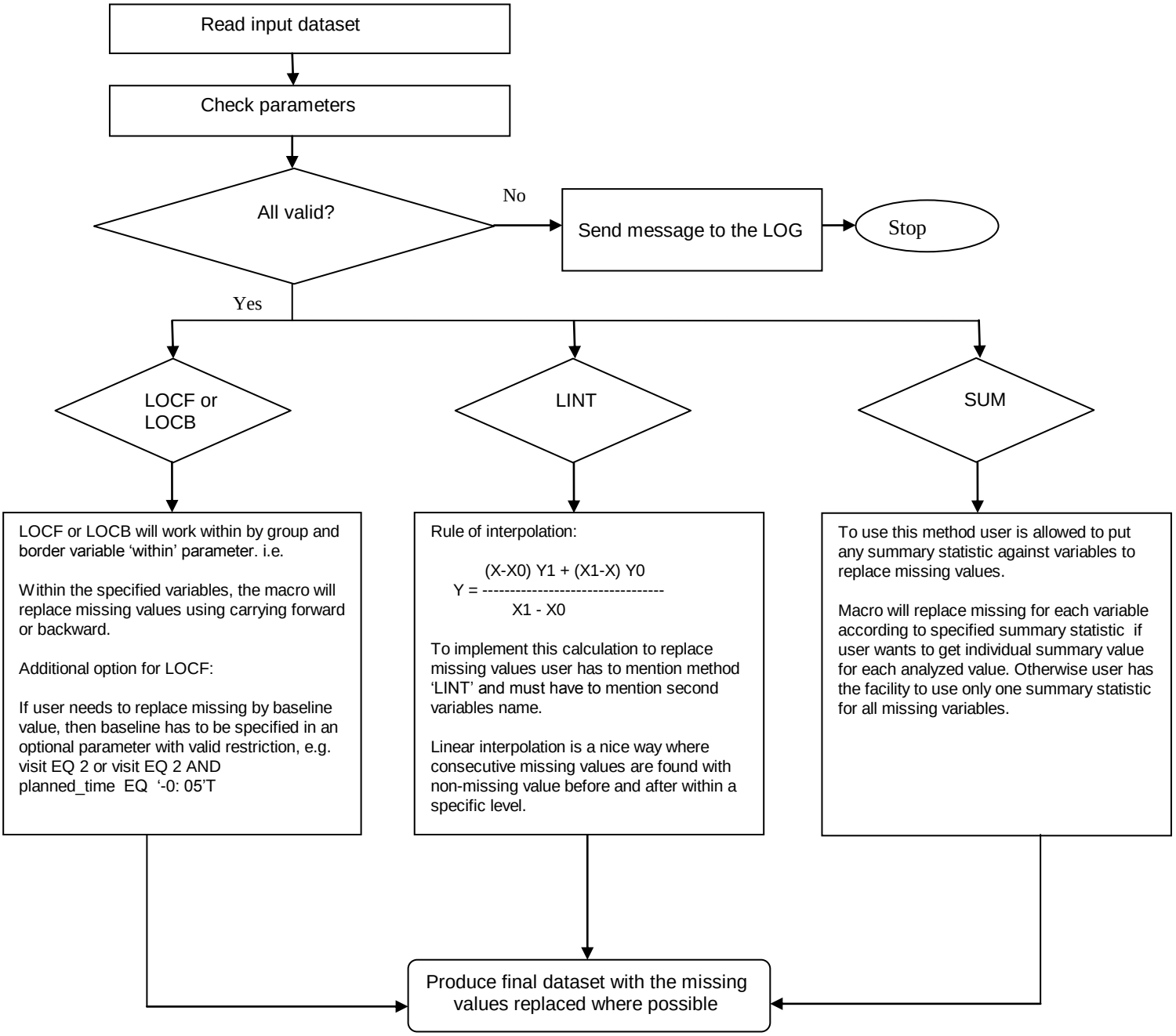
SUM : To replace missing by summary values at some level are required to use this method.

LINT : Applying linear interpolation procedures to irregularly-sampled raw data can obtain time series with equidistant sampling intervals. The application of approximation methods to the time series produces function we can use to fill in gaps in the data under stationary conditions.

Summary of parameters:

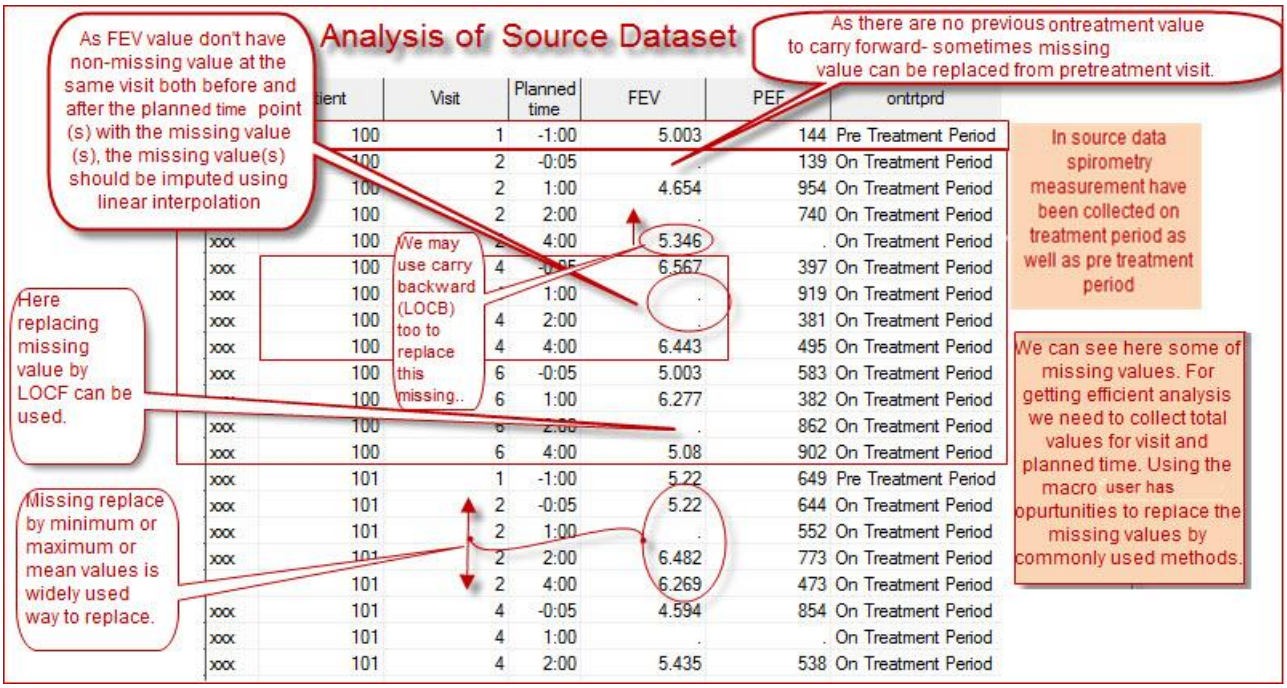
- Input and output dataset names.
- A variable or a list of variables which should be checked for missing variables.
- Order and grouped variables needed to specify how values will be carried forward or backward to get previous or next values for different calculations. All method must use this grouping condition. Valid values are one variable or a list of variables separated by space.
- Methods used to replace missing values. These are LOCF (Last Observation Carried Forward), LOCB (Last Observation Carried Backward), SUM (replace missing by summary values (i.e.- Mean, Max, Median etc) and LINT (Linear Interpolation). There is a special option for LOCF to replace missing values by baseline or pre-treatment values.
- Summary statistic to use (i.e. Mean or Median or Max etc.) when SUM method is selected. User can replace all imputed variables by a single summary value, or specify a method for each variable which is checked.
- Define baseline condition (e.g. visit EQ 2 or visit EQ 2 AND planned_time EQ '- 0:05'T).
- To replace consecutive missing with non-missing before and after value in any level, user must specify one valid variable name in right place. User must specify the second variable with all known values (current, previous and next values) to calculate interpolation. Valid value can be any one those are being used in group variables.
- To get an extra option from LOCF to replace missing by pre-treatment values (this is not as usual case), user has to mention specifically and this way values will be replaced by pre treatment value if there are not available value for carrying forward. This also can be done by using LOCB if pre –treatment values are not allowed for specific study.

MACRO PROCESSING:



EXAMPLE MACRO CALL:

Source dataset with missing values:

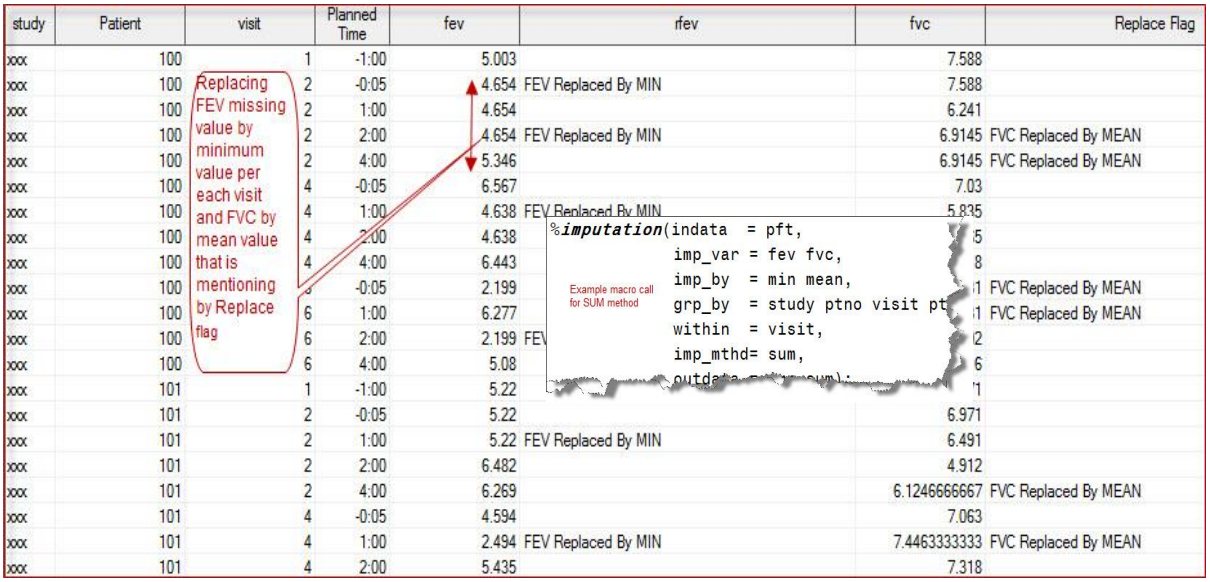


ient	Visit	Planned time	FEV	PEF	ontrtprd
100	1	-1:00	5.003	144	Pre Treatment Period
100	2	-0:05	.	139	On Treatment Period
100	2	1:00	4.654	954	On Treatment Period
100	2	2:00	.	740	On Treatment Period
100	2	4:00	5.346	.	On Treatment Period
100	4	-0:05	6.567	397	On Treatment Period
100	.	1:00	.	919	On Treatment Period
100	4	2:00	.	381	On Treatment Period
100	4	4:00	6.443	495	On Treatment Period
100	6	-0:05	5.003	583	On Treatment Period
100	6	1:00	6.277	382	On Treatment Period
100	6	2:00	.	862	On Treatment Period
100	6	4:00	5.08	902	On Treatment Period
101	1	-1:00	5.22	649	Pre Treatment Period
101	2	-0:05	5.22	644	On Treatment Period
101	2	1:00	.	552	On Treatment Period
101	2	2:00	6.482	773	On Treatment Period
101	2	4:00	6.269	473	On Treatment Period
101	4	-0:05	4.594	854	On Treatment Period
101	4	1:00	.	.	On Treatment Period
101	4	2:00	5.435	538	On Treatment Period

Macro calls using different methods and its output example:

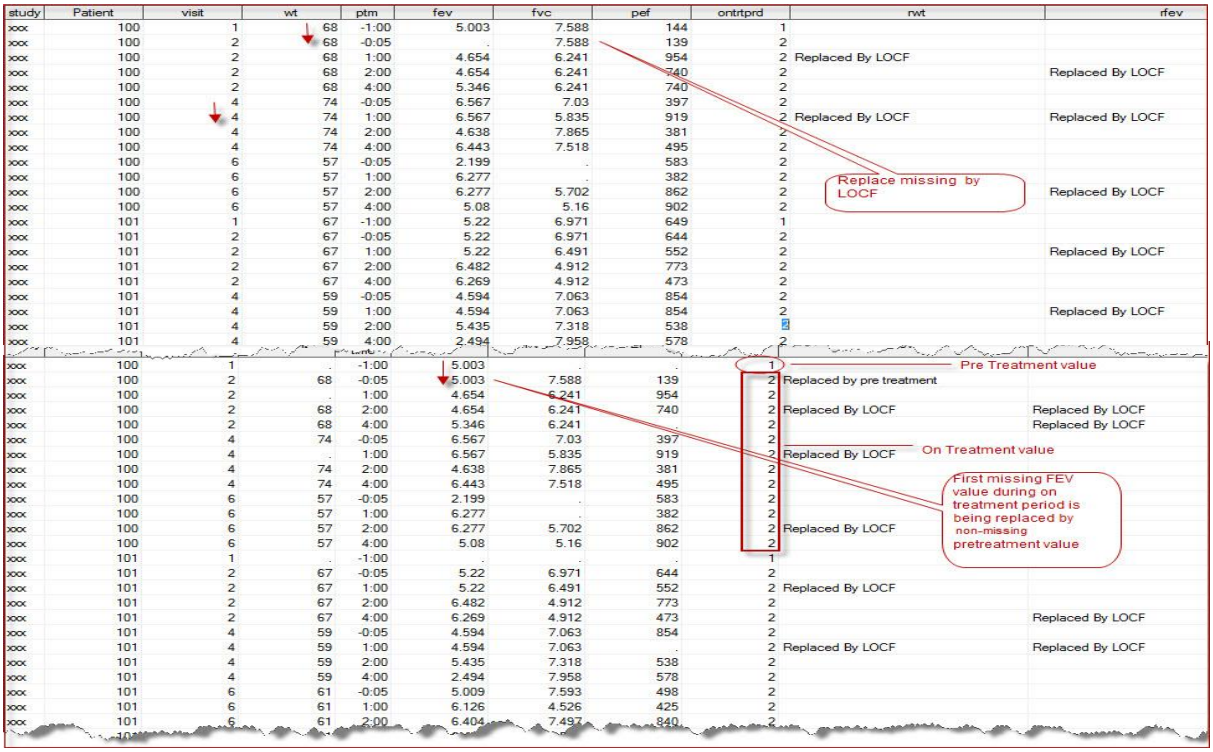
a. Missing replaced by SUM method:

SUM method is widely used way to replace missing. To get the right output user has to put the required parameters. Output will be the following after replacing missing values.



study	Patient	visit	Planned Time	fev	rfev	fvc	Replace Flag
xxx	100	1	-1:00	5.003		7.588	
xxx	100	2	-0:05	4.654	FEV Replaced By MIN	7.588	
xxx	100	2	1:00	4.654		6.241	
xxx	100	2	2:00	4.654	FEV Replaced By MIN	6.9145	FVC Replaced By MEAN
xxx	100	2	4:00	5.346		6.9145	FVC Replaced By MEAN
xxx	100	4	-0:05	6.567		7.03	
xxx	100	4	1:00	4.638	FEV Replaced By MIN	5.835	
xxx	100	4	2:00	4.638		5	
xxx	100	4	4:00	6.443		8	
xxx	100	5	-0:05	2.199		1	FVC Replaced By MEAN
xxx	100	6	1:00	6.277		1	FVC Replaced By MEAN
xxx	100	6	2:00	2.199	FEV	2	
xxx	100	6	4:00	5.08		6	
xxx	101	1	-1:00	5.22		1	
xxx	101	2	-0:05	5.22		6.971	
xxx	101	2	1:00	5.22	FEV Replaced By MIN	6.491	
xxx	101	2	2:00	6.482		4.912	
xxx	101	2	4:00	6.269		6.1246666667	FVC Replaced By MEAN
xxx	101	4	-0:05	4.594		7.063	
xxx	101	4	1:00	2.494	FEV Replaced By MIN	7.4463333333	FVC Replaced By MEAN
xxx	101	4	2:00	5.435		7.318	

b. Missing replaced by LOCF/LOCB and special LOCF by forwarding non-missing value from pre-treatment values using an additional option.



c. Missing replaced by baseline value.

Study	Patient	Visit	Planned time	FEV	rfev	FVC	rfvc
xxx	100	1	-1:00	5.003		7.588	
xxx	100	2	-0:05	.		7.588	Baseline value
xxx	100	2	1:00	4.654		6.241	
xxx	100	2	2:00	.		7.588	Replaced By baseline value
xxx	100	2	4:00	5.346		7.588	Replaced By baseline value
xxx	100	4	-0:05	6.567		7.03	
xxx	100	4	1:00	.		5.835	
xxx	100	4	2:00	4.638		7.865	
xxx	100	4	4:00	6.443		7.518	
xxx	100	6	-0:05	2.199		7.588	Replaced By baseline value
xxx	100	6	1:00	6.277		7.588	Replaced By baseline value
xxx	100	6	2:00	.		5.702	
xxx	100	6	4:00	5.08		5.16	
xxx	101	1	-1:00	5.22		6.971	
xxx	101	2	-0:05	5.22		6.971	
xxx	101	2	1:00	5.22	Replaced By baseline value	6.491	
xxx	101	2	2:00	6.482		4.912	
xxx	101	2	4:00	6.269		6.971	Replaced By baseline value
xxx	101	4	-0:05	4.594		7.063	
xxx	101	4	1:00	5.22	Replaced By baseline value	6.971	Replaced By baseline value
xxx	101	4	2:00	5.435		7.318	
xxx	101	4	4:00	2.494		7.958	
xxx	101	6	-0:05	5.009		7.593	
xxx	101	6	1:00	6.126		4.526	
xxx	101	6	2:00	6.404		7.497	
xxx	101	6	4:00	5.22	Replaced By baseline value	4.576	

Visit 2 and -0:05 planned time being considerd as baseline here. Missing values are being replaced non-missing baseline values.

d. Missing replaced by using Linear Interpolation (LINT)

Study	Patient	Visit	Planned time	FEV	rfev	PEF	fvc	rfvc
xxx	100	4	-0:05	6.567		397	.	
xxx	100	4	1:00	6.50252	Replaced By LINT	919	.	
xxx	100	4	2:00	6.5256666667	Replaced By LINT	381	.	
xxx	100	4	4:00	6.443		495	.	
xxx	100	6	-0:05	5.003		583	.	
xxx	100	6	1:00	6.277		382	.	
xxx	100	6	2:00	5.878	Replaced By LINT	862	.	
xxx	100	6	4:00	5.08		902	.	
xxx	101	1	-1:00	5.22		649	.	
xxx	101	2	-0:05	5.22		644	.	
xxx	101	2	1:00	5.87624	Replaced By LINT	552	.	
xxx	101	2	2:00	6.482		773	.	
xxx	101	2	4:00	6.269		473	.	
xxx	101	4	-0:05	4.594		854	.	
xxx	101	4	1:00	5.03132	Replaced By LINT	689.68	.	
xxx	101	4	2:00	5.435		538	.	
xxx	101	4	4:00	2.494		578	.	
xxx	101	6	-0:05	5.009		498	.	
xxx	101	6	1:00	6.126		425	.	
xxx	101	6	2:00	6.404		840	.	
xxx	101	6	4:00			999	.	
xxx	102	1	-1:00	4.943		462	.	
xxx	102	2	-0:05	4.943		457	.	
xxx	102	2	1:00	5.55868	Replaced By LINT	655.64	.	
xxx	102	2	2:00	6.127		584.33333333	.	
xxx	102	2	4:00	6.81		839	.	
xxx	102	4	-0:05	5.99		729	.	
xxx	102	4	1:00	6.799		834	.	
xxx	102	4	2:00	6.154	Replaced By LINT	803.33333333	.	

Two values were missing and values were available before and after both planned time within the visit. So linear interpolation is being used.

LIMITATION OF THE MACRO

a) Either all the variables which are being replaced use the same imputation method, or the method have to be explicitly specified for all variables which are being replaced
b) The structure of the data is fixed, so updates to the macro will be required if the data structure changes. It means that one macro to cover all studies is not possible, but it is possible to use one macro for a project containing many studies.

CONCLUSION

The macro is easy to use, ensures consistency within trials, and when used on a project level, it ensures consistency across studies within a project. For a small organization like us, it is versatile enough to use with various clients with no or minimal change, thus saving us a great deal of time. Having one macro which does this also means that everyone is familiar with the macro, and is therefore comfortable to use it.

ACKNOWLEDGMENTS

Special thanks to Mizan Alam , Rafi Rahi and Aminul Islam for testing this macro and investigating different approaches to solve issues with missing values.

CONTACT INFORMATION

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

Kamrunnisa
Shafi Consultancy Bangladesh
50/B, Borobazar, Amberkhana
Sylhet, Bangladesh
Phone: +88 01928098077
E-mail: kamrunnisa@shaficonsultancy.com
Web: www.shaficonsultancy.com

Mokhfur Alam Chowdhury
Shafi Consultancy Bangladesh
50/B, Borobazar, Amberkhana
Sylhet, Bangladesh
Phone: +88 01730049205
Email: mokhfur@shaficonsultancy.com
Web: www.shaficonsultancy.com