

PHUSE 2012

Macro to replace missing values

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 Forward (LOCF), Last Observation Carried Backwards (LOCB), Linear Interpolation and using summary statistics 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 implement these different methods for replacing missing values. The macro is designed to be used with standard data, and it helps to both reduce programming and validation time, as well as ensuring consistency both within studies and across studies. The macro can be adapted with minor changes for different dataset structures and for different methods, 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 macros to replace missing values.

- **Macro aims:**

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.

- **Problems encountered and how they were resolved:**

There were two main aims of the macro, to be easy to use for replacing missing values, and to be flexible enough for different types of data. To ensure the macro is easy to use meant reducing the number of macro parameters to the bare essentials. This means the users are clear what each parameter they are specifying do, making them more in control and feeling less like they are using a black box. To allow the macro to be flexible for different types of data was more of a challenge. Macro parameters can be used to define which method is used to replace which variable, but to make this easy to define for users is a challenge. Making the macro call more like a procedure is the solution. As users are used to calling procedures, making the macro call more like a procedure helps to make the call more user friendly, easy to remember and more importantly easy to use.

- **Limitations 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 for use 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.