murdoch children's research institute **Version: 1.1**	**Title:** **Good Stata Programming**

Murdoch Children's Research Institute

Authors: Katherine Lee and Francesca Orsini

Date Created: 9 December 2011

Date reviewed: 15 April 2021

Approved by: Katherine Lee

Date: 19 April 2021

Revised for MISCH & Approved by: Julie Simpson, Director of MISCH HubDate: 15/01/2021

Change History

9 December 2011: Initial version created

15 April 2021: Initial version reviewed, cosmetic changes made

Contents

1. PURPOSE

To describe the steps and procedures to be followed for writing Stata programs in order to prepare for and perform a statistical analysis. To describe the steps for maintenance of a controlled working environment via a folder structure system

2. APPLICABILITY

This procedure applies to all studies for which responsibility for the statistical analysis rests with a CEBU statistician.

3. ROLES AND RESPONSIBILITIES

Statistician: responsible for importing datasets into Stata (which may or may not be from a different environment), checking the internal consistency and validity of the data, and generating datasets for analysis including creating derived variables defined in the protocol or the Statistical Analysis Plan (SAP). The statistician is also responsible for the development of Stata programs to perform the analysis of the study data according to the protocol or SAP, including the production of the required tables and figures.

4. DEFINITIONS

Statistical Analysis Plan (SAP): a comprehensive and detailed description of the methods of data analyses proposed for study (most often used in randomised controlled trials).

Database lock: once data management activities (data cleaning, queries resolution, serious adverse event reconciliation and coding activities) have been performed and there are no pending issues the database is considered locked and it cannot be changed (again often only used in randomised controlled trials).

Unblinding: when a member of the study team requests to know the treatment allocation of participants in the study (randomised controlled trials only).

5. INTRODUCTION

Stata is a statistical analysis package which CEBU uses for the following activities:

- Database import from a different environment (i.e. MS Access, EpiData, MS Excel...etc), if required
- Checking the internal consistency and validity of collected data (to the extent possible without referring to source data or clinical knowledge)
- Generation of datasets for analysis, including the merging of datasets, changing of data structure from wide to long format or vice versa calculation of derived variables
- Generation of results for reports, including tables and figures

5.1 Components of a Stata session

There are three components in a Stata session: the *input* file, the *program* file and the *output* file.

- The *input* file used by Stata is a data file with suffix *.dta* which has been obtained directly from the study database or a converted form of the study dataset from another computer program (generally either the original study database or an excel file).

- Stata *programs* are called do-files and act upon the *input* file (identified by the suffix *.do)*. The do-file contains Stata commands that the statistician wishes to execute. Executing a do-file is equivalent to executing a series of commands interactively, but carries out all of the commands in one go. By collecting the required commands in a do-file allows the statistician to quickly reproduce work performed previously (for example after the input file has been updated). Note do-files may contain reference to ado-files or "automatically loaded do-files". An ado-file is similar to a do-file but unlike do-files they do not need to be referred to in their specific location in order to run the file. When you type a command that Stata does not know, it automatically looks for an ado-file of that name in the Stata program files on your computer. If Stata finds an ado-file with the required name, Stata automatically loads and executes it as with any other command built into Stata. You can also ask Stata to look in other specific folders on your computer for ado-files (see Section 5.2).

- The *output* file is the storing of what appears on screen when you run your do-file. These files have the suffix *.log* (which stores the data in ASCII text format, or *.smcl* for use within Stata only), and for this reason they are usually called log-files. At the beginning of every do-file the statistician should ask Stata to open a log-file by typing log using filename. The log-file captures all of the commands and the output from the command from the time of opening the log-file. After typing log close the corresponding log-file will be ready for viewing and printing of the results generated by the execution of the do-file. Additionally, at the start of every do-file before typing log using filename include the command capture log close, which will close a log if any is open and do nothing if no log is open (see Section 5.2).

Note there are other outputs that can be obtained from Stata e.g. graphs, new datasets, tables…, but these outputs are specific to the analysis being carried out, and not part of the general set-up within Stata.

5.2 Do-file structure: overview

Each Stata do-file should be structured to meet the following criteria:

- Well commented
- Easily understandable
- Purpose, input requirements and output produced clearly defined
- Written focusing on a specific content (i.e. demography, medical history, primary efficacy, etc...)
- Linked to a log-file

It is recommended that you start each do-file with the following commands:

```
capture log close
version XX.X
clear all
```

```
set more off
log using \log_files\filename.log, text replace
```

This clears the workspace, frees up memory to speed the calculations, and opens a Stata log (be sure to end the do-file with log close). The first do-file you should set up for any project is a Master do-file (see Section 5.2). This do-file, (among other things) sets up your working directory – which we recommend to be the core folder for the date specific analysis for details. All other do-files should then use paths which are relative to this core folder.

Do-file names. The name of each do-file should reflect its content. All do-files written to generate analysis datasets should start with cr while all analysis files should start with an. It may be useful to name create files as cr1_X, cr2_Y,… and analysis files an1_X, an2_Y, or an1_tab_1, an2_fig_1,… as this can help yourself and others to know which order the do-files should be run in. It is also advisable that the log-files are named with the same name as the do-file that generates them.

Variable names. Variable names should reflect what data the variable represent in the dataset. Ambiguous names should be avoided. Variable names ideally should be short, all alphabetical letters lower case, so that they are easy to retype. It is best to avoid capital letters in variable names as Stata is case sensitive. The use of variable labels is recommended to provide additional details on the content of the variable. Value labels should also be used where applicable. When referring to a variable within a do-file it is advisable to avoid using abbreviated variable names.

Comments. Putting a * at the beginning of a line in the do-file tells Stata not to execute that line. This serves to annotate the do-file. Commenting is an essential part of the program, as it facilitates understanding and will assist others in navigating your do-files. A detailed description of the steps reported in the program should be included within the code. In particular, comments should report any useful and not obvious information.

Header. Each do-file should have a comment block at the top of the do-file to describe the purpose of the do-file, who generated it and when, and to provide any useful information regarding what the do-file does.

As a minimum it is recommended that the header includes the following:

- The study identifier
- The name of the program
- The author and the date of creation
- The purpose
- Notes about program updates (when program is updated)

For example:

```
* Study:          Project\Analysis X
* Program name:   Table_X.do
* Creation date:  dd-mm-yyyy
* Purpose:        e.g. Create Demography Table
* Author:         Name Surname
* Version:        YYYYMMDD
* Note:           XXXXX
* Update:         XXXXX, dd-mm-yyyy, FirstName Surname
```

6. PROCEDURE

6.1 Project structure

Each project should have its own electronic folder. In order for other statisticians in the unit to be able to navigate the folder it is important that all project folders have the same structure.

When starting a new project the statistician should copy the project folder template stored in \CEBU statistics and data management\CEBU Guidelines and SAPS and paste it in their work directory as new study file which should be renamed to reflect the name of the project. The structure of the project folder is reported in the Appendix 9.1: it is recommended that the statistician follows this basic structure for all projects, especially for the *Analyses* sub-folder. Other sub-folders may be personalised based on the project characteristics and peculiarities, in particular the Data Management sub-folder which may need to be adapted depending on the project.

In particular it is suggested that:

- A new folder should be opened for each project
- Each project should have a separate folder which follows the same template in order to have a standardized and structured fashion which will be easily sharable
- A sub-folder should be opened for each analysis within the project – Data_Analysis_and_Reporting_A, Data_Analysis_and_Reporting_B, etc.
- This sub-folder should contain all of the data, do-files, output and reports from the specific analysis for the project.

Once a statistical report has been written and finalised for a project, the sub-folder for that analysis should be closed and no further changes should be made to the datasets, the do-files and the log-files within the sub-folder. If additional results are required for the project after this, additional lines of code or additional do-files can be added to do-files within the folder, and a new report produced. However, if additional analyses are required which change the output from the analysis already presented, a separate folder within the analysis folder, labelled with a new date (new sub-folder *YYYYMMDD*) should be started.

If a number of changes are made to the analysis and/or report during the life of a project, it may be useful to keep a master list of the changes and additional analyses required for the project, along with the reasons for the changes made.

6.2 Master do-file

The first program to be generated for a new project should be the Master do-file. This do-file is used for:

- Setting the working directory for the project: we recommend this to be the core folder for the date specific analysis e.g. Project_Folder\Analyses\Analysis_X\YYYYMMDD, where Analysis_X is used to distinguish between different analyses within the same study e.g. analysis of the main study vs a sub-study. These subfolders may or may not be needed. All commands within the do-files that read or write files will then use paths which are relative to this main folder. This means that if the project folder is moved for some reason, only this line of code requires amending and all of the do-files will still run.

- Setting the memory for the project
- Adding additional paths for ado-files (if required). Stata looks for ado-files in the Stata program files on your computer. However it may be that you want to write your own commands and store them on your personal directory in a place where they can be used in a number of projects. In order for you to simply refer to the command that you have written you need to tell Stata where to look for your user-written commands. This can be achieved using the adopath command in Stata. The alternative is to put ado-files in the current study directory, although this means that the command is available only when the statistician works in that directory.
- Listing all of the do-files related to the analysis in the order that they should be run to carry out the entire analysis. It can be useful to comment out this section of the do-file so that the Master do-file can be run at the start of a session working on the project without necessarily running all of the do-files associated with the project.

See Appendix 9.2 for a template of a Master do-file. Master do-file should be stored in the directory Project_Folder \Analyses\Analysis_X\YYYYMMDD\Do_files (see Appendix 9.1).

The Master do-file must always be executed when starting a session working on the project before running any other do-file.

6.3 Data import

Once the statistician receives the raw data, the dataset(s) should be stored in the specific study folder under Project_Folder\Analyses\Analysis_X\YYYYMMDD\Data\Source_Data (see details in Appendix 9.1). These raw data should be stored in their original format. No changes should be applied to these datasets.

If not in Stata format, the statistician should import the dataset(s) into the Stata environment by generating and using a specific Stata do-file that executes an import procedure, using the insheet, import or infile command as part of the create do-file (see section 5.5).

6.4 Randomisation list import

In case of a randomised controlled trial (RCT) the statistician should request a copy of the randomisation list. Once received, the statistician should store the randomisation list in the study folder Project_Folder\Analyses\Analysis_X\YYYYMMDD\Data\Source_data (see details in Appendix 9.1) in its original format. If not in Stata format, the statistician should import the randomisation list into Stata as with any other input file. No changes should be applied to the randomisation codes.

Unblinding of RCTs **-** If the study is a blinded RCT, once the randomisation list has been obtained, the statistician should sign and date the code break form. Once completed, this form should be stored in the dedicated study file which should be kept by the Principal Investigator.

6.5 Creation of analysis datasets

For the purpose of analysis, the statistician should create analysis datasets using Stata do-files (according to the general principles described in Section 4.1 above) using the raw data sent to CEBU as input. These “create” do-files should be stored in the do-file directory Project_Folder \Analyses\Analysis_X\YYYYMMDD\Do_files (see Appendix 9.1), and, as suggested, should have the suffix “cr”. The output from a create do-file should be a cleaned dataset with all the variables required for a particular analysis. When generating a create do-file the following principles should be followed:

- There should be one do-file per analysis dataset
- It can be useful to add a number after the cr suffix in the do-file name in order to know which order they should be run
- A do-file may merge more than one raw dataset in creating a new analysis dataset
- Derived variables should be generated according to the algorithm defined in the protocol or SAP
- Variable and value labels should be used where required
- The output from a create do-file should be a cleaned dataset that is ready for analysis. Each created dataset should be saved in the folder Project_Folder \Analyses\Analysis_X\YYYYMMDD\Data\AD using a save statement
- Every do-file should begin by opening a log-file which should be stored in the folder Project_Folder \Analyses\Analysis_X\YYYYMMDD\Log_files. When the do-file is written and processed the statistician should visually check that the log-file is error-free.

6.6 Statistical analysis programming

In order to obtain statistical results (e.g. a set of tables and/or figures), the statistician creates a set of “analysis” Stata do-files (according to the general principles described in Section 4.1). Data manipulation should be avoided within these analysis programs: these files should use the analysis datasets generated in the create do-files as input and should not make any changes to the dataset (all changes/additions to the dataset should be made in the create do-files). Each do-file name should start with the suffix an and reflect its content. A sequential number can be added after the an suffix in order to know which order analysis do-files should be run. These analysis do-files should be stored in the do-file directory Project_Folder \Analyses\Analysis_X\YYYYMMDD\Do_files (see Appendix 9.1).

The statistician should write all the necessary do-files in order to perform the analysis as specified in the protocol or SAP by the investigator. This may include:

- Creation of tables
- Creation of figures
- Analysis of study results

The analysis should generally be carried out using a separate do-file for each table and each figure, although there are circumstances where one do-file may be used to produce multiple tables/figure, for example if producing a number of idenitical figures for a range of laboratory measurements.

Tables generation. Ideally the tables required for an analysis should be summarised in a List of Tables and analysis should be carried out using a separate do-file for each table. Each of these do-files should generate both a log-file (as every do-file should) and, for do-files that generate a table, a dataset which contains the results for the required table: in this way there will be as many do-files as tables/figures. It is recommended that the name of the do-file contains details of the table/figure to which the program refers (e.g. an1_tab_1). The tables of results (i.e. summary statistics, frequency tables and model estimates...), should be exported from Stata as Excel files, named with the same title as the do-file itself, both of which should reflect the table that it is being generated. Exported files should be saved within the Project_Folder \Analyses\Analysis_X\YYYYMMDD\Results\Tables directory (see Appendix 9.1). In creating the output datasets/tables it is recommended to use commands like postfile, collapse or contract to collect the required data in a (new) Stata dataset, followed by the export excel command to export the tabulated data to an Excel format. Once exported, the table dataset can be opened in Excel. The multiple tables can then be sent to researchers directly, or can be copied and pasted into a single Word document, which will constitute the Tables & Figures Document for the analysis.

In cases where listings of patients' data are required, the listings required should be summarized as a List of listings, and the same procedure described above should be followed for the generation of the patients' data listings document.

Figure generation. As for the generation of tables, the generation of the figures required for an analysis should ideally be carried out with a separate do-file for each figure. Again, each of these do-files should generate a log-file. To save the figures in a format that can be used outside of Stata, for example to be inserted into a Word document, it is advisable to use the graph export command and save it as either Portable Network Graphics files (.png), which have the advantage that they can be viewed on any operating system, or Windows Enhanced Metafile (.emf) or TIFF (.tif). If the do-file generates a single figure, it can be useful to save the figure with the same title as the do-file for consistency. The generated figure should be saved within the Project_Folder\Analyses\Analysis_X\YYYYMMDD\Results\Graphs directory (see Appendix 9.1).

Once the statistical analysis has been completed and the tables and the figures have been produced, they should be presented, along with commentary, in the statistical report.

7. TIMING

Statistical activities requiring Stata programming should not be started until after the approval of the protocol and the SAP and the database lock if applicable (i.e. formal randomised controlled trials). The Code Break Form used to obtain permission to break the blinding, when required, must be approved before the randomisation code is obtained and data analysis is performed.

8. APPENDICES

Appendix 9.1. – Project Folder

- **Documents**
    - Protocol
    - References
    - Statistics
        - *SAP*
        - *Sample size*
        - *Other*
- **Data Management***
    - Database
- **Protocol**
  References
  Safety review or DMC
  Statistics
- **Analyses**
    - **Analysis_X**[+]
        - YYYYMMDD
            - Data
                - *Source_data*
                - *AD*[#]
            - Do_files
            - Log_files
            - Results
                - *Statistical_Report*
                - *Paper*
                - *Tables*
                - *Graphs*

[#]: AD stands for Analysis Datasets

[+]: The title of the sub-folder should be chosen to reflect the analysis required, for example, main analysis vs analysis of a sub-study. This may not be required if there are not multiple analyses for a given study. If use these names should be reasonably short.

* No structure is recommended for the data management sub-folder as this will depend on the needs of the study.

Note a separate YYYYMMDD folder should be started for each analysis of a project. If anything changes in the results aside from the addition of new analyses, a new folder YYYYMMDD should be established within the project folder. Once the analysis is finalised the YYYYMMDD folder should be closed and no further changes made to do-files (e.g. set to read-only).

Appendix 9.2 – Master.do Template

```
**************************************************
*       Study:  Project\Analysis X
*       Program Name:  Master
*       Creation Date:  dd-mm-yyyy
*       Purpose:  MASTER programme
*       Author:  Name Surname
*       Version:  X
*       Note:  XXXXX
*       Update:  XXXXX, dd-mm-yyyy, Name Surname
**************************************************

clear all
version XX.X
set more off

cd "X:\Study\Project_Folder\Analyses\Analysis_X\YYYYMMDD"

/*

* Create do-files

do “Do_files\cr1_predictor.do”
do “Do_files\cr2_demography.do”
do “Do_files\cr3_outcome.do  ”
...


* Analysis dofiles

do “Do_files\an1_tab_1.do”
do “Do_files\an2_tab_2.do”
do “Do_files\an3_tab_3.do”
...

*/

exit
```

Appendix 9.3 – General Do-file template

```
******************************************
*         Study:  STUDYNAME
*         Program Name:  an_table_1.do
*         Creation Date:  dd-mm-yyyy
*         Purpose:  Table 1
*         Author:  Name Surname
*         Version:  01
******************************************

clear all
set more off


* Opening log
log using “Log_file\an_table_1.log", text replace

* Import analysis dataset Demo
use "Data\AD\Demo.dta", clear

.
.
.
commands
commands
commands
.
.
.

* Save table 1 as an excel file
export excel using "Results\Tables\table_1.xls”

log close

exit
```