**Statistical Monitoring Program Instructions.**

**6 Inliers**

The function **inlier\_check** is used to investigate data sets with several continuous variables. It looks for participants who have several values which like close to the means.

Parameters to give the function:

1. data

This is a data frame with the patient id in the first column, the site name/number (must both be numeric) in the second column and any continuous variables to check in the following columns.

Data frames can be read in with the following code:

**options(stringsAsFactors = FALSE)**

**data.rand<-data.frame(read.table("STUDY12\_REG.txt", row.names=NULL, header=TRUE, sep="\t"))**

(This would read in a text file called STUDY1*2\_REG.txt* and store it in the data frame *data*.*rand*)

1. min.part

This is the minimum number of participants a site can have and still be included – any site with fewer than min.part participants will be excluded (a list of sites which are excluded are output).

This MUST be at least 2 (though 5 or more is suggested). If a value of 1 or 0 is entered this is automatically replaced with 2 (a message is output to say that this has been done).

1. trial.name

The name of the trial. This will be used to label the output files. For example:

**trial.name<- “STUDY12”**

1. n

Any patient that falls more than n SDs below mean will be selected as an inlier.

1. pervar

Pervar is logical and should be set as TRUE if you wish to calculate the average distance from the mean (i.e. the sum of the distances divided by the number of variables used to calculate it) or FALSE if not. If a participant has lots of missing values then the sum of distances from the mean may be smaller than a participant whose individual points lie closer but who has no missing data. Dividing the sum by the number of non-missing values can eliminate this problem.

**Calling the function**

Once the program and the parameters above are stored in R’s memory, the program can be run using the following command:

**inlier\_check(data, min.part, trial.name, n, pervar)**

Where each parameter is stored as in 1-5

**The output:**

The program outputs 2-3 text files and a plot for each site (with enough participants).

**The plots**

Each site with at least min.part participants is plotted. The plot shows log(*d*) (see article for more details) against the participant index). Any participant who lies more than *n*SDs (of log(*d*)) below the mean log(*d*) is circled in red, these are the inliers.

The example below shows inlier plots for sites 11-21 in the registration data in study 12. This plot has a name in the form: “*INLIERS\_2.5\_SDs\_STUDY12\_REG\_2014-07-22\_sites\_11-21.wmf*”, where STUDY\_12 was the specified as the trial.name, REG as the sheet.name, n was given as 2.5, and 22/07/2014 was the date the program was run.



Sites are plotted in groups of 9, ordered by site number, as above.

**The text files**



Excluded sites

The first text file is only created if sites are excluded (due to having too few participants).

Right, an example of the first text file. This has a name in the form: *“INLIERS\_excluded\_sites\_STUDY12\_REG\_new\_2014-07-22.txt”* where trial.name was specified as STUDY12 and the date (22/07/2014) was the date the program was run.

 Information about the inliers.

The second text file gives more detailed information about the inliers found (see example below, left). If no inliers are found it contains the message “No inliers found”).





This has a name of the form: “*INLIERS\_2.5\_SDs\_STUDY12\_REG \_2014-07-22.txt*”

Information about the data checked.

The third text file (example, right) gives information about all of the data checked.

This has a name of the form:

*“INLIERS\_DATA\_TESTED\_STUDY12\_REG \_2014-07-22.txt”*

**Warnings:**

With the exception of values of min.part given as less than 2, there are no error messages coded into the function. If data is not read in as above, the function may not work as it should, or possibly at all. Please take care when creating the parameters from your data.