**Statistical Monitoring Program Instructions.**

**8. Correlation Check**

The function **correlation\_check** can be used on continuous variables to see if the correlation structure differs between sites.

Parameters to give the function:

1. data:

This should be in the form of a data frame with the site number in the first column (must be numeric – if a string is given please recode as numeric) followed by the continuous measurements to check (i.e. columns 2+).

Data frames can be read in with the following code:

**options(stringsAsFactors = FALSE)**

**entry.data<-data.frame(read.table("ABC02\_entry.txt", row.names=NULL, header=TRUE, sep="\t"))**

(This would read in a text file called *ABC02\_entry.txt* and store it in the data frame *entry.data*.)

1. trial.name

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

**trial.name<- “ABC02”**

1. sheet.name

The name of the CRF being checked, again this will be used to label the output files. For example:

**sheet.name<- “entry”**

1. min.obs

This the minimum number of observations/patients a site can have and still be checked (we used 10; the program will may not run (or may get stuck at the simulation stage) if very small numbers are chosen).

**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:

**correlation\_check(data, trial.name, sheet.name, min.obs)**

Where each parameter is stored as in 1-4

**The output:**



**The plots**

There is one grey scale plot for each site (with enough observations to be assessed). This show the correlations represented as a series of grey scale squares. Correlations of -1 will be white, +1 will be black and everything else a shade of grey in-between. The function puts 9 plots onto each plotting display (this could be altered – see comments in the code).

 An example of 4 grey scale plots can be seen right.

This plot has the file name:

*“Correlation\_ABC02\_entry\_2014-07-14plot\_sites\_6-31.wmf”*

Where ABC02 is the trial name (this is whatever has been specified as trial.name when calling the function), entry is the name of the CRF (specified as sheet.name) and the date would be replaced with the date the program was run.

Sites 6-31 give the first and last site numbers plotted (not all sites shown above).

**The text files**

Two text files are output. The first has a name in the form: “CORRELATION\_FORMAL\_TEST\_ABC02\_entry\_2014-07-14”This gives information from the simulations; the site numbers tested, the d value, the number of patients and the p-value.



The second text file has a name in the form *“CORRELATION\_DATA\_TESTED\_ABC02\_entry\_2014-07-14”* this gives information about the data that has been tested:



**Warnings:**

There are no error messages coded into the function (except the built in R errors described below). 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.

The program will remove all variables with very small SD values (<1e-05) as these may result in an NA correlation, however it only checks the SDs for each variable in the original data (i.e. the dataset with all patients). If there is a small/zero SD in one or more variables in a site then this will stop the program and produce an error message, this variable or site must be removed.

The problem described above can also occur when running the formal test if R happens to randomly select a group of patients with a very small/zero SD in one or more variables. If this happens a warning message will be produced but the program will not stop – it will simply skip this random site and generate a new one.

Running the formal tests can take a while so R keeps track of the site being tested (click on the console as it runs and it will update) so you can tell the program is still running correctly.