## Requirements specification document - porting Stata mimix program to R.

## Background to mimix

-mimix- performs reference based multiple imputation for sensitivity analysis of a longitudinal clinical trial with protocol deviation.

Specific qualitative treatment arm based assumptions for missing data are made.

### Specifying imputation method

The five imputation options of mimix are randomized-arm MAR, jump to reference, last mean carried forward, copy increments in reference, and copy reference. Each method constructs appropriate joint distributions for the observed and unobserved data for deviating individuals based

on an underlying assumption for the missing data. These joint distributions imply conditional distributions for the missing data, given the observed, from which imputations can be drawn.

### Algorithm for obtaining imputed data sets

mimix implements the general algorithm of Carpenter, Roger, and Kenward (2013)

1. Separately for each treatment arm, take all the observed data, assume MAR, and fit an MVN distribution with an unstructured mean ( a separate mean for each of the baseline and the postrandomization observation times) and a variance-covariance matrix using a Bayesian approach

with an improper prior for the mean and an uninformative Jeffrey’s prior for the covariance matrix.

2. Draw a mean vector and covariance matrix from the posterior distribution for each treatment arm, we use the MCMC method to draw from the appropriate Bayesian posterior, with a sufficient burn-in, and we update the chain sufficiently in between to ensure that subsequent draws are independent. The sampler is initiated using the expectation-maximization (EM) algorithm.

3. Use the draws in step 2 to form the joint distribution for each deviating individual's observed and missing outcome data as required, depending on the assumption for the missing data.

4. Construct the conditional distribution of missing given observed outcome data for each individual who deviated from the joint distribution formed in step 3. Sample missing data from the conditional distributions to create a completed dataset.

5. Repeat steps 2-4 m times, resulting in m imputed datasets.

Steps 1 and 2 are implemented by Stata’s mi impute mvn command with mcmconly option.

e.g. mi impute mvn `mi\_impute' , mcmconly burnin(`burninM') priorVIM(jeffreys) initmcmc(em, iter(1000)) saveptrace(`mimix\_parms\_a`i'', replace)

mi impute mvn fills in missing values of one or more continuous variables using multivariate normal

regression. It accommodates arbitrary missing-value patterns. You can perform separate imputationson different subsets of the data by specifying the by() option. mi impute mvn uses an iterative Markov chain Monte Carlo (MCMC) method to impute missing values

Uses data augmentation (DA), iterative MCMC assuming multivariate normal model.

### Analysis

### The analysis on the imputed datasets are either regress or mixed

### For each of the imputed data sets either linear regression models of the dependent variable on treatvar (at final timepoint) and any included covariates, or linear mixed models using REML are run and results combined using Rubin’s rules.

### Data structure

(equivalent to Stata long format)

**Inputs**

Response/dependent variable: Name of response variable

ID: Name of variable identifying the subject(individual)

Time: Name of the variable identifying units of time

Treat: Name of the variable identifying the (treatment) Group

m: Number of imputations

method (or methodvar) Which imputation method out of 5 to use

Covariates Covariates must be constant over time and none missing

refgroup or refgroupvar level of treat var chosen for reference group, methods j2r,cir,cr

interim specifies imputation method for all interim missing values

iref specifies level of treat var chosen as reference for interim

Mixed or regress saturated repeated-measures model, (uses Stata mi estimate),

using REML or linear regression of depvar on treatvar at final t

Note before using mi impute mvn must have mi set the data, also mi register the independent vars

Clustering is accounted by including the factor variable, i.response

R packages mice,Amelia,pan,mi,norm,norm2,BaBooN,VIM

Comments from mimix program

\*! mimix v1.4 23apr2018

\*PRE-PROCESSING AND ERROR CHECKING

\*NOTE: refgroupvar methodvar mimix\_refgroupvar ARE TEMPVAR's CREATED LATER ON

\*SAVE THE ORIGINAL DATA

\*SAVE THE DATA THAT WILL NOT BE SENT INTO MATA TO ADD TO IMPUTED DATA AT END

\*PROCESS COVARIATES: TOKENIZE THEN CHECK COMPLETENESS

\*methodvar & method ARE MUTUALLY EXCLUSIVE OPTIONS:

\*ONE OF methodvar or method MUST BE SPECIFIED

\*ORDER DATA HERE IN TERMS OF COVARIATES ORDER:

\*EXIT IF NO TREATMENT VARIABLE (OR RESPONSE) HAS BEEN SPECIFIED

\*refgroupvar & refgroup ARE MUTUALLY EXCLUSIVE OPTIONS:

\*CHECK refgroup IS A VALID NUMBER IF ITS A NUMBER, i.e. A TREATMENT LEVEL IF THIS OPTION IS USED (STRING PICKED UP LATER)

\*CHECK method SPECIFICATION

\*CHECK interim SPECIFICATION

\*CHECK iref IS A VALID NUMBER OR STRING, i.e. A TREATMENT LEVEL IF THIS OPTION IS USED

\*IF INTERIM IS CR, J2R of CIR check WE HAVE A iref SPECIFICAITON

\*CHECK m IS >=1

\*REMOVE OBSERVATIONS WITH INCOMPLETE COVARIATES & DISPLAY THE NUMBER DROPPED.

\*CHECK COVARIATES ARE THE SAME ACROSS TIME POINTS BY ID NUMBER. IF NOT WARN AND TAKE FIRST VALUE.

\*DROP VARIABLES WE DON'T NEED IN THE IMPUTATION MODEL

\*CHECK ONE ENTRY PER ID PER TIME POINT - REQUIRED - IF NOT ADVISE APPROPRIATE ACTION

\*SAVE ORIGINAL DATA - WITH ANY CORRECTIONS MADE FOR COVARIATES

\*FOR LATER RECODE:

\*RECODE iref

\*RESHAPE

\*ALSO CHECKS IMPUTATION METHOD IS CONSISTENT WITHIN ID AND TREATMENT GROUP/RESPONE AROUND CORRECT WAY

\*CREATE MISSING DATA PATTERN VARIABLE

\*ADD INTERIM VALUE INDICATOR IF TREATING INTERIM VALUES DIFFERENTLY

\*IF interim - CREATE A VARIABLE THAT HOLDS THE NAMES OF THE INTERIM MISSING VALUES. SAVE THIS VARIABLE AND `id' FOR LATER DATA MERGING

\*RECODE `id' AS A NUMERICAL VARIABLE IF NOT THE CASE FOR SUBSEQUENT MATRIX MANIPULATION OF THE DATA. SAVE RECODE FOR LATER MERGING.

\*VARIABLES TO BE IMPUTED IN mi impute mvn

\*CHECK AND RECODE THE REFERENCE GROUP VARIABLE TO CORRESPOND WITH `mimix\_treat'

\* RUN MI IMPUTE MVN

\*MI SET THE NEW IMPUTED DATA SET

\*REGRESSION OPTION

\*MIXED OPTION

\*SAVING AND CLEAR OPTIONS

\*RETURN OF iref

\*DROP MATRICES NO LONGER REQURIED