**README FOR CHAPTER 7 CODE**

Table of Contents

[General information 1](#_Toc36475174)

[Packages 1](#_Toc36475175)

[Data information 2](#_Toc36475176)

[Program information 3](#_Toc36475177)

**DISCLAIMER:** The data supplied is simulated at random from the real distributions of the respective variables, the code will run and produce results and figures, but the results are meaningless

# General information

* All programs run from a root/parent directory, where all the provided files and directories should be stored. At the top of each R program, set the working directory to this, or place an Rproject in this root directory and then nothing needs to be added to the code.

* The programs take raw datasets with missing data, do the imputation and then the analyses. The simulated datasets which the programs are run on are stored in the /data/CSV2019/ directory. More details about these are in the ‘Data information’ section.
* There is a batch file (AA\_run\_all\_batch.sh) that will run all the programs in order to carry out the analysis from the UNIX command line. However the working directory must be set at the top of each .R file in order to do this. To use the .Rproj approach, I recommended running the programs in the Rstudio gui.
* The figures that will be produced when running the code are already provided in the /figures/ directory. Code can be re-run to compare.
* Programs were run using R.3.4.2.

# Packages

The package versions in the below table were used. These were the package versions installed on the remote computer I used for the analyses so some are quite old, however in most cases I see no reasons why the latest package versions wouldn’t work.

One exception is the mice package. I have found when using the most recent version (3.8.0), it is not possible to extract the padded dataset containing dummy variables. These are necessary as I specifically call on dummy variables categorical variables to create interaction terms with age during the imputation process. Therefore I recommend installing mice version 2.46.0. I have provided code to do this:

install.packages("devtools")

require(devtools)

devtools:install\_version("mice", version = "2.46.0")

But have often found various issues when using devtools on different machines, this is the only bit of code I am not confident will run smoothly on your computer so may require some detective work from the user to get the correct version installe.d

|  |  |
| --- | --- |
| Package | Version |
| mice | 2.46.0 |
| foreach | 1.4.4 |
| doParallel | 1.0.11 |
| tidyverse | 1.3.0 |
| ggpubr | 0.1.8 |
| knitr | 1.20 |
| survival | 2.42-3 |

# Data information

I provide information about the variables in the datasets, in case you want to repeat the analysis using your own data

**1) Allpers**: This contains all patients that were eligible for inclusion in the primary prevention cohort (eligible for statins for primary prevention). All date variables are in the format ‘01/01/1990’, apart from first cvd event, which is in form ‘01/01/90’. The variables are:

Patid

Gender: (1 = male, 2 = female)

Dtvalid: start of valid follow up

Dtcens: end of valid follow up

Do85: date turned 85

first\_statin: date of first statin (if no statin, this is set to 19/02/2042)

first\_cvd\_all: date of first cvd event (if no CVD event, this is set to 19/02/42)

index\_date\_A: study start date/date of entry into primary prevention cohort, latest of aged 25 + 1 year valid follow up, or 1st Jan 1998

**2) statin\_users\_analysis\_dataset:** This contains all patients who were initiated on a statin before the end of follow up, or before having a CVD event. Variables needed to calculate QRISK3 CVD risk score are measured at the time of statin initiation. I will not list all variables here, refer to the fake data files to see variable names.

All binary variables are 0 = absent, 1 = present.

All continuous variables are just the number, no transformation

Smoking = 0 (never smoker), 1 (ex-smoker), 2 (current smoker), and is then edited for Numcigs\_per\_day

Townsend = 1 (least deprived), 2, 3, 4, 5 (most deprived)

Qrisk\_score = qrisk score, if recorded in the database

# Program information

The table below details what is done in each program.

|  |  |
| --- | --- |
| Program | Definition |
| 0.1 | Install packages |
| 1.1 | Produce baseline table of statin users dataset |
| 1.2 | Impute the statin users dataset 10 times |
| 1.3 | Impute the statin users dataset another 10 times (20 imputed datasets in total now) |
| 1.4 | Combine the imputed datasets into one workspace |
| 2.1 | Load functions to calculate QRISK3 scores for patients in the statin users dataset (to calculate EHR derived risks) |
| 2.2 | Calculate EHR derived risks for patients in the statin users cohort |
| 2.3 | Run analysis to calculate incidence of statin initiation each year |
| 2.4 | Run all other analyses:  1) Calculate average risk of patients initiated on statins each year  2) Calculate proportion of patients in each risk category initiated on statins each year  3) Check agreement between EHR derived risk scores and coded risk scores |