**README FOR CHAPTER 3 CODE**

Contents

[General information 1](#_Toc36475955)

[Packages 1](#_Toc36475956)

[Data information 2](#_Toc36475957)

[Program information 4](#_Toc36475958)

**DISCLAIMER:** Simulated datasets to run this code on have not been provided. The code is provided for transparency.

# General information

* The code files, running from file p1.1.1 to p6.4 are designed to be run in order. Although many can be run simultaneously, a few are reliant on earlier programs. This code assumes the existence of various objects: the imputed analysis datasets themselves, a copy of the original analysis dataset with data missing and a cohort of statin users with statin use dates already derived.
* Unlike the other chapters, everything does not run relative to a parent directory as I was a less experienced coder at the time. This means that in order to re-use this code, each program must be searched for points at which various directories containing relevant datasets must be directed to. Detailed instructions for this are found in the *Data information* section,
* I state what must be included in these files and where they must be located for the programs to run. This code could be adapted for use on any dataset, not just one derived from CPRD.
* All analyses have been run in R version 3.4.1

# 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 |
| Survcomp | 1.28.5 |
| survAUC | 1.0-5 |
| Hmisc | 4.1-1 |
| CPE | 1.4.4 |
| pec | 2018.04.18 |
| reshape2 | 1.4.2 |

# Data information

*Data locations*

There are four key locations where .RData files and csv datasets are loaded from and stored. In the programs these are referred to as A, B, C and D

For A, B and C please search every file for “setwd( “ and change to a relevant directory on your system. For D please search for “*read.table*” and change to a relevant directory on your system

Directory A must contain four .RData files which contained the 20 imputed datasets for female and male cohorts respectively. These are called

* *imputed\_datasets\_loaded\_female.RData*
* *imputed\_datasets\_loaded\_male.RData*
* *imputed\_datasets\_loaded\_female\_C.RData*
* *imputed\_datasets\_loaded\_male\_C.RData*

In these .RData files the 20 imputed datasets must be in long format and stored in a list called “*long\_data\_parallel”*. The first two files must contain the imputed datasets from the primary cohort used for model development and validation. The third and fourth (suffix \_C) contain the imputed datasets of the cohort of patients registered in CPRD in 1st Jan 2016. We only actually use the first one of the imputed datasets, but this is how the data is set up. In each of these .RData files there must also be an object of the original analysis dataset with missing data, called “*anal.dat*”, this was fed into the imputation procedure. This means any transformations to variables in “*long\_data\_parallel*” are also present in “*anal.dat*”. This dataset should have also been sub-setted to the correct gender.

The variables names within each long dataset should be assigned based on their names in the models. The only one we believe isn’t clear, is family history of CVD, which is referred to as Famhis\_lstrict. The rest should be clear, e.g. Atrialfib = Atrial Fibrilation.

Directory B is where is where all output (.RData files) is stored.

Directory C is where the region file is stored. The region file contains a data frame with two columns, “pracid” and “region”. This gives the respective region of each practice, which is used for models E and F. This data frame is stored in a file called “*region.dat.RData*”.

Directory D is one final directory which contains csv files. This contains the raw datasets that have not been imputed for male and female cohorts or had any transformations applied to them: development cohort and 2016 cohort. It also contains the statin users cohort, which has a row for every statin treatment period (patients may be in this multiple times). This file also contains the 10 year risk, as calculated by model B, at the start of the statin treatment period. All these datasets contain data for both genders. The required variables in the statin cohort are explained in the R programs. These files are always read in directly using the read.table command.

Again you must check in whether the “*read.table*” command is used, and change the directory to the appropriate one.

The names of the files in this directory are:

The statin users cohort: *statin\_users\_cohort.csv*

The development cohort: *analysis\_dataset\_A\_comb.csv*

The 2016 cohort: *analysis\_dataset\_C\_comb.csv*

**Key information when fitting the models**

When imputing the data, we centred age. This ended up being problematic at after calculating optimal fractional polynomials, age had to be rooted at zero (or close to it, 0.05). This means adding “18.1” to age in each female development cohort, and adding “16.9” to age in each male development cohort. Assuming you have made no transformations to age, you can simply deduct 24.97 from it to get the appropriate transformation.

When using programs p3.1 - p3.4, and calculating risks for the 2016 cohort, age should be adjusted accordingly. This means deducting the **same** amount off age that was deducted from the development cohort, for consistency across the two cohorts. Given we had also centred age when imputing the 20106 cohort, this meant we had to add back on the mean age of the 2016 cohort, before deducting the mean age of the development cohort and then adding on either “18.1” or “16.9”. In hindsight, this was very frustrating and we regret centring age during imputation.

If following these instructions, hopefully you do not have to repeat this process, and this part of the code can be removed, and instead just make a simple adjustment to age (deduct 24.97).

Note that the code is frequently commented to make it easier to understand.

# Program information

|  |  |
| --- | --- |
| **File name** | **Information about group** |
| **P1.1.1** | **Derive risks for female cohort according for model A – F.** |
| **P1.1.2** |
| **P1.1.3** |
| **P1.1.4** |
| **P1.1.5** |
| **P1.1.6** |
| **P1.1.7** |
| **P1.2.1** | **Derive risks for male cohort according for model A – F.** |
| **P1.2.2** |
| **P1.2.3** |
| **P1.2.4** |
| **P1.2.5** |
| **P1.2.6** |
| **P1.2.7** |
| **P1.3.1** | **Combines risks and creates tables for primary analysis. Dependent on all P1.1 and P1.2 files being successfully run.** |
| **P1.3.2** |
| **P1.3.3** |
| **P1.3.4** |
| **P2.1.1** | **Calculates: CH, ρk, R2 ,D, R2D and CGH for models A, B, C and D, and CH, ,D, R2D and for models E and , for the female cohort.** |
| **P2.1.2** |
| **P2.1.3** |
| **P2.1.4** |
| **P2.1.5** |
| **P2.1.6** |
| **P2.2.1** | **Calculates: CH, ρk, R2 ,D, R2D and CGH for models A, B, C and D, and CH, ,D, R2D and for models E and , for the female cohort.** |
| **P2.2.2** |
| **P2.2.3** |
| **P2.2.4** |
| **P2.2.5** |
| **P2.2.6** |
| **P2.3.1** | **Calculates R2PM for the male and female cohorts respectively** |
| **P2.3.2** |
| **P2.3.3** |
| **P2.3.4** |
| **P2.4.1** | **Calculates ρw,a for the male and female cohorts respectively** |
| **P2.4.2** |
| **P2.4.3** |
| **P2.4.4** |
| **P2.5.1** | **Calculates CU for the male and female cohorts respectively** |
| **P2.5.2** |
| **P2.5.3** |
| **P2.5.4** |
| **P2.6.1** | **Calculates IBS and R2IBS for the male and female cohorts respectively** |
| **P2.6.2** |
| **P2.6.3** |
| **P2.6.4** |
| **P2.6.5** |
| **P2.6.6** |
| **P2.6.7** |
| **P2.6.8** |
| **P3.1** | **Calculates risks for patients registered on 1st Jan 2016 according to models A - F** |
| **P3.2** |
| **P3.3** |
| **P3.4** |
| **P3.5** | **Combines risks for patients registered on 1st Jan 2016 and produces Tables and Figures** |
| **P3.6** |
| **P3.7** |
| **P3.8** |
| **P4.1** | **Takes development cohort and statin users cohort and calculates crude secular trends, and also fits poisson models to model incidence** |
| **P4.2** |
| **P4.3** |
| **P4.4** |
| **P5.1** | **Calculates hazard ratios from model B and model E, as well as code for calibration of model B** |
| **P5.2** |
| **P5.3** |
| **P5.4** |
| **P5.6** |
| **P5.7** |
| **P5.8** |
| **P6.1** | **Code for producing Figures 1 – 4.** |
| **P6.2** |
| **P6.3** |
| **P6.4** |