Assessment 1 - Statistical Modelling and Inference for Health

Dr Glen Martin and Dr Matthew Sperrin

This assessment aims to test your understanding of section 1 (Modelling complex data) of the Statistical Modelling and Inference for Health module. You are required to undertake a data analysis (guided by the indicated questions) of the provided datasets and report your findings. Please format your answers into a report-style and interpret the statistical output. While you are encouraged to provide your R code to document your working, we do **not** expect to simply see R console output copied-and-pasted: it is the interpretation and explanation of your analysis that is needed alongside this. For example, you might like to structure your report with section heading being the questions, and your written answers the body of each section with any code output formatted into tables/figures (with captions).

Assignment settup

For this question, we will be using a dataset called "Framingham". The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of participants in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. Participants have been examined biannually since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. For this assignment, we will be exploring an anonymous subset/extract of these data.

For this assessment, we will restrict to the first observation for each participant. Therefore, the data includes 4434 patients. After setting your working directory appropriately, load the dataset into R (note the use of tidyverse package):

```r
library(tidyverse)
frmgham_data <- read_rds("frmgham_data_firstobs.RData")
```

The assignment is split into two parts: in part A, we are interested in investigating whether use of anti-hypertensive medication at exam time (variable "BPMEDS") is associated with time-to-cardiovascular-disease (CVD); in part B, we want to study whether we can develop a prediction model to predict time-to-death during follow-up.

Part A

1) The primary outcome is time-to-developing cardiovascular risk (CVD). Given this outcome, please write down the considerations we need to make in terms of analysis choice (hint: think carefully in terms of other outcomes that are recorded that could prevent someone developing CVD)? (5 marks)

2) Summarise the data with appropriate exploratory analysis/plots. (5 marks)

3) Using a non-competing risk analysis (i.e. using the outcome variables TIMECVD and CVD), fit relevant Cox proportional hazards model(s) to test the association between use of anti-hypertensive medication at exam time with time-to-CVD. For every Cox model that you fit, remember to test the proportional hazards assumption. What can we conclude regarding our primary question of whether use of anti-hypertensive medication at exam time is associated with time-to-CVD? (15 marks) What are the implications for competing risks? (5 marks)

Part B

In this part of the assessment, we are interested in developing a prediction model to predict risk of death during follow-up.

There is missing data in some predictors; therefore, start this part by running the following code to perform a complete case analysis (we will explore alternative methods later in the course!):

```r
frmgham_data_complete <- frmgham_data %>%
  filter(complete.cases(.))
```

4) Using the complete data above (frmgham_data_complete) to fit a Cox proportional hazards model where the outcome is time-to-death, with appropriate predictor variables. Report the model in an appropriate way, which would allow someone to use the model to make predictions. (10 marks)

5) When evaluating the predictive performance of this model on the same data as was used to develop it, we find it has excellent calibration and a C-statistic of 0.75 (discrimination). Write a short paragraph explaining whether such results are informative in terms of the internal validity of the model/ (5 marks)

6) Upon external validation of the model in a new setting/population, we found the performance of the model to be insufficient. Write a short paragraph explaining what steps could be done to address this. (5 marks)