Grand Challenges Assignment 1

A look into predicting future myocardial events based on different physiological factors

This assignment will be looking into one of the current methods, the Framingham model, used by health professionals to determine an individual's five-year cardiovascular risk and assessing the significance of the different parameters. My research question is "Are the current parameters used to assess one's cardiovascular risk sufficient or is there more that can be investigated"? This data analysis was motivated due to the sheer number of Australians who are affected by heart disease every year. In 2020, heart disease would claim one life every eighteen minutes, was the leading cause of death in males and was responsible for nearly one in five deaths (17.1%) [1]. As with many medical conditions, the negative outcomes of cardiovascular disease can be mitigated if action is taken early. However, this can sometimes prove to be difficult as many individuals will report no signs or symptoms of comorbid health conditions such as high blood pressure or high cholesterol.

Aforementioned, the Framingham is one of the current methods used by health professionals and requires several parameters to determine its output. These are an individual's gender, diabetic status, smoking status, age, systolic blood pressure and total cholesterol as seen in appendix 1. My work will investigate to see if the Framingham model is an adequate prediction tool for cardiovascular risk. My prediction for this analysis is that the current method will yield statistically significant results reinforcing its place in medical practice.

The dataset used to help determine the research question was sources from Kaggle (link in references). It contained results from 303 participants in which their;

- Age
- Sex
- Type of chest pain (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)
- Resting systolic blood pressure (mmHg)
- Total cholesterol (mg/dL)
- Diabetic status (1 = true, 0 = false)
- ECG result (0 = normal, 1 = ST-T wave normality, 2 = left ventricular hypertrophy)
- Maximum heart rate achieved from stress test (bpm)
- ST wave depression induced by exercise relative to rest, or oldpeak (
- The slope of the line between the last two variables
- Number of heart vessels obstructed
- Thall (not clearly specified what this was indicative of so was removed)
- Exercise induced angina status (1 = yes, 0 = no)
- Outcome variable (1 = high chance of cardiovascular event or >50% narrowing of artery, 0 = low chance of cardiovascular event or <50% narrowing of artery)

Were recorded. As I was unable to find any concrete mention of what the "Thall" column represented it was removed from the dataset for my analysis.

This dataset was chosen as it contained 5 out of the 6 parameters used for the Framingham model with 'smoking status' being the only one missing. It also had an easily manipulated outcome variable which would make producing graphs and statistical findings less challenging. Many of the remaining columns of data were of interest as well since they could potentially provide additional sources or test parameters that could be included in the prediction of future cardiovascular events.

Firstly, to help understand the dataset, each column was allocated to a new sub array. These sub arrays were called continuous, categorical, and outcome. All columns that had a continuous input such as age or blood pressure were placed into the first sub array, categorical columns such as diabetic status or chest pain status were placed into the second sub array and the output variable was placed into the outcome array. This was done to help better visual potential relationships that could be interpolated from the dataset between the variables. After this the cholesterol reading was converted from mg/dL (American standard units) to mmol/L (Australian standard units) to better relate the findings to that of the Framingham model which uses mmol/L.
Next, using the seaborn library in python the data was visualised in different types of graphs. Countplots were produced for the categorical data to help visualise the total tally and proportion of the data for each individual parameter. Boxenplots were used for the continuous data to easily display key statistical findings such as mean and quartile distribution.

To help answer my research question "Are the current parameters used to assess one's cardiovascular risk sufficient or is there more that can be investigated" the 5 parameters that my chosen data set and the Framingham model had in common were first selected for graphical and statistical analysis. Once again, the seaborn library was used to produce kdeplots (see appendix 3). I chose this style of graph as I was able to divide my independent variable into two separate area curves based on the target outcome and display it all on a single set of axes. The overall mean was also inserted onto the graphs to give some reference to any potential skew for either area graph. This allowed me to easily to interoperate results and findings at first glance. To help prove the significance of any finding, a two-sample t-test was performed on the continuous variables. I would compare to see if there was a negative t stat value with a p value of <0.05. The benefit of performing this analysis was to reject the null hypotheses.

The following hypothesis were drawn from the Framingham model.

1) If a person is male, then they will be more likely to have an outcome variable of 1 (high risk) because more males die from cardiovascular disease compared to females.
2) If a person has diabetes, then they will be more likely to have an outcome variable of 1 (high risk) because diabetes is a comorbid health condition that increases one's risk of cardiovascular complications
3) If a person is older, then they will be more likely to have an outcome variable of 1 (high risk) because as an individual ages, they are more likely to experience cardiovascular complications
4) If a person has higher systolic blood pressure, then they will be more likely to have an outcome variable of 1 (high risk) because as one's blood pressure increases, their heart is placed under more stress.
5) If a person has higher total cholesterol, then they will be more likely to have an outcome variable of 1 (high risk) because as one's total cholesterol increases, they are

more likely to develop plaque deposits on the arteries leading to narrowing of the vessel.

The first hypothesis was invalid. This was due to the ratio of gender to outcome variable for males and females being 0.8:1 and 3:1 respectively. This was the opposite of what was predicted in the hypothesis as seen in appendix 3.

The second hypothesis was invalid. Once again, the ratio of diabetic status to outcome variable was not of prediction. For diabetics, the ratio was 1:1 and non-diabetics was 1.2:1 as seen in appendix 3.

For the third hypothesis, the null hypothesis was rejected with a p value of $3.75x10^{-5}$ ($p<0.05$). However, the younger a person was had a statistically significant undesirable outcome on their cardiovascular health with the mean age of the high risk group being 52.5 years and the low risk group being 56.6 years as seen in appendix 3. Meaning the third hypothesis was invalid

For the fourth hypothesis, the null hypothesis was rejected with a p value of 0.00577 ($p<0.05$). However, with lower blood pressure had a statistically significant negative outcome on their cardiovascular health. The mean blood pressure of the low risk group was 134.4 mmHg and for the high risk group 129.3 mmHg as seen in appendix 3. Meaning the fourth hypothesis was invalid.

The fifth and final hypothesis accepted the null hypothesis with a p value of 0.0693 ($p>0.05$) indicating that there was no significant difference for an individual's outcome variable based on their total cholesterol as seen in appendix 3. Meaning the fifth hypothesis was invalid.

As all five of my initial hypotheses were rejected, I went back over the data and started looking for other tools that were not included in the Framingham model that may be good indicators for cardiovascular risk. I did a similar analysis on the maximum heart rate achieved during a simple stress test and the outcome variable. This yielded the first positive result. When assessing if an individual with a higher maximum heartrate was more likely to have a high risk outcome, the null hypothesis was rejected with a p value of $8.49x10^{-15}$ ($p<0.05$) and mean values of 139.1 and 158.5 bpm for low risk and high risk respectively.

Lastly, when interpreting the results for the different types of chest pain compared to target the outcome, I found a very predicable result. The majority of the low risk individuals were classified as having normal/stable angina where for the high risk individuals there was a slight increase as the categories increased from stable angina to unstable angina. Based on the physiology of this condition and its related comorbidities with future cardiovascular events this was expected.

In conclusion, The Framingham test's appeal comes from its ease of use and cost effectiveness, some other methods used, such as angiograms, can be costly to the governing body and take up other valuable resources. I believe that adding a parameter such as maximum heart rate after a stress test would be very fitting in the current framework and appeal of the Framingham test and potentially help identify a few more at risk individuals than currently. I also believe that even though my initial hypotheses were all rejected, these parameters should remain in the Framingham test. A major limitation that I faced when performing this data analysis was the sample size of the study used. The appropriate sample size for this to be relevant to Australia would be roughly 5113 participants compared to my 303.

On final notes, I have high hopes that health professionals will be able to lower the impact of cardiovascular risk on the Australian population as technology evolves and we discover new methods for detecting health conditions. I also believe that this is an interesting space for potential artificial intelligence work to be done as that is currently an emerging specialty for the field.

References

[1] Heartfoundation.org.au. 2022. *Key Statistics: Heart Disease | The Heart Foundation*. [online] Available at: <https://www.heartfoundation.org.au/bundles/for-professionals/australia-heart-disease-statistics> [Accessed 9 September 2022].

[2] Hua, X., McDermott, R., Lung, T., Wenitong, M., Tran-Duy, A., Li, M. and Clarke, P., 2017. Validation and recalibration of the Framingham cardiovascular disease risk models in an Australian Indigenous cohort. *European Journal of Preventive Cardiology*, 24(15), pp.1660-1669.

[3] Kaggle.com. 2022. *Heart Failure Prediction with EDA & 4 Models*. [online] Available at: <https://www.kaggle.com/code/oykuer/heart-failure-prediction-with-eda-4-models/data> [Accessed 9 September 2022].

Appendix

As I was unwell for the proposal presentation, I presented just to my class tutor Mohsi. So, although I was unable to receive feedback from my peers, I noted down some of his feedback and that is what I will be addressing for this section. It should also be noted that I changed topic from my proposal to my final presentation.

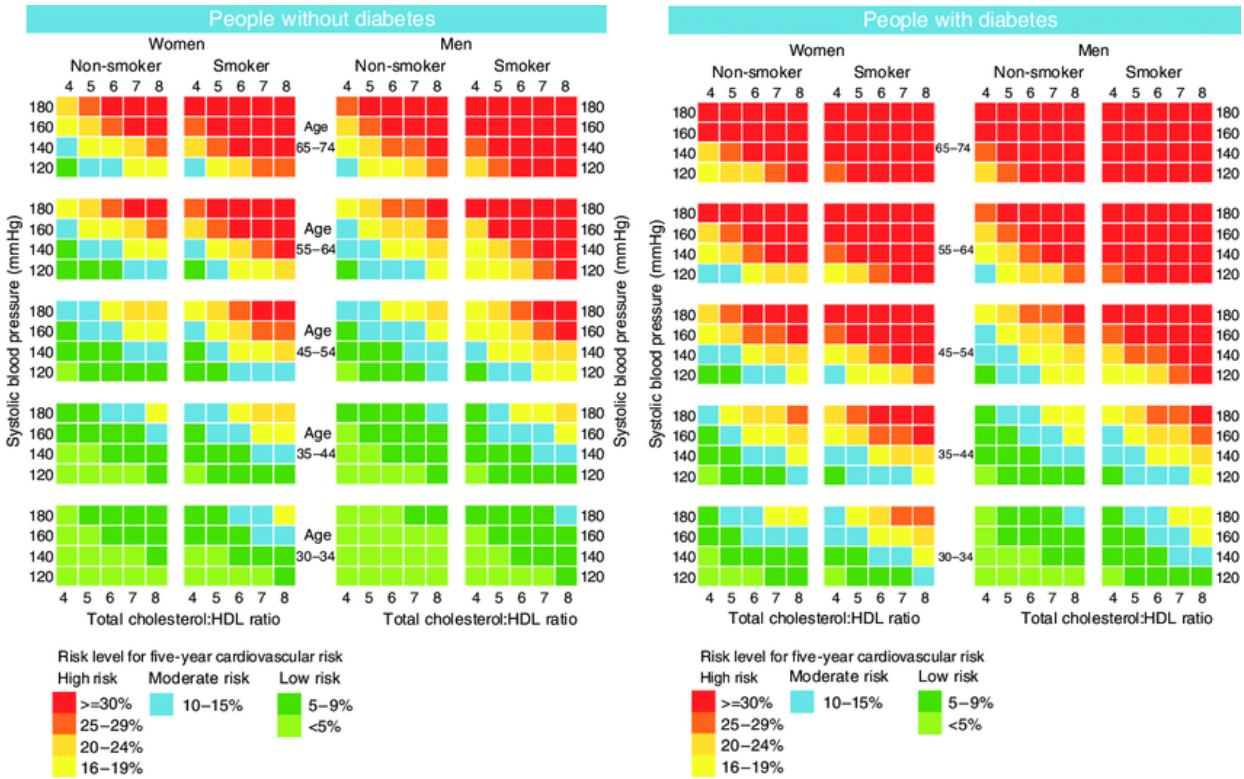
My first point was regarding how I had worded my research question. It was in the format of is *independent variable* causing an increase in *dependant variable*. It was explained to me that it would form a better research question if it was in the reverse structure. So, when forming my research question for my new topic I tried to word my question in the most coherent way possible.

I also did not mention how I was going to address my research question. Once again, I reflected upon this advice and tried to clearly explain my methods that I will be using to help answer my question.

My final point of feedback was that my strategy should be to look at current research papers and see what has already been investigated in the field. It was for this reason I based my question around potentially finding new parameters to add or consider in the Framingham model.

To summarise, I hope I was able to positively reflect on the feedback from my tutor and improve my work for the final report.

Appendix 1: the Framingham model [2]



Appendix 3: kdeplots showing relationship between independent variables and the outcome variable