Introduction:

This report will explore a cardiovascular disease dataset. This dataset consists of many demographic and physical factors of a sample of a group of patients who had been referred after having complained of heart-related pain/discomfort. The data itself was collected before and after patients were asked to perform 10 minutes of static activity on an exercise bike. With a sample size of just over 300 observations for each variable and no missing data, the sample itself is quite small considering the deadliness of CVD worldwide. However, based on the results of this investigation, this report will attempt to determine the following question: What factors, both demographic and physical, are predictive of the presence of cardiovascular disease in patients experiencing chest discomfort?

words:122

Exploratory Analysis

From the initial look at the data and boxplots, it seemed that age, sex and fasting glucose levels would be less successful predictors of CVD. In all cases, the three predictors showed little variation between groups of patients who had CVD and those who did not. Age having similar means and ranges across sufferers and non-sufferers suggested that beyond knowing the general age range of sufferers was between the mid-40s to the mid-60s, we could not say much more about the predictor’s effect on CVD.

Employment, somewhat unsurprisingly, showed that the skill level of labour has an inverse proportion with CVD, as those working in manual labour and low-skill professions experienced the greatest amount of sufferers. This was a similar case for the highest education level reached by patients, as the raw data clearly shows that those who left school earlier in their lives had a higher within-group proportion of CVD sufferers.

Chest pain was a predictor that showed a relationship between the level of angina found in patients and CVD, with those having atypical angina recording the most CVD patients. Resting beats per minute (BPS) and Cholesterol levels also seemed to show that those with CVD had a much tighter interquartile range of values taken, suggesting possibly that CVD restricting normal function can be viewed in reverse to be predictive of CVD itself. However, from the values themselves, there was no apparent difference in the means across both groups of the outcome variable.

Other predictors that showed early signs of significance were Thalass, a blood disorder that restricts haemoglobin function, thought to be related to CVD. Here, the level of interest was the fixed defect level, where the within-group percentage jumped. This suggests that those with fixed thalassemia defects are at a much greater risk of CVD than those with reversible defects or none.

words:305

First to FInal Model:

	Model 1	Final Model
Age		N/A
Sex		N/A
Employment	***	***
Edu_level		N/A
Chest_pain		*
Rest_bps		N/A
Chol		N/A
Fgl		N/A
Rest_ecg	*	**
Max_r		N/A
Ex_ang	*	**
ST_diff	*	**
Slope		N/A
Diab_num	**	***
Thalass	*	*
Residual Deviance/Null Deviance	113.35/417.6	89.78/417.638

Groupings of variables

Demographic	Physical
Age	Chest_pain
Sex	rest_bps
Employment	chol
Edu_level	fgl
-	rest_ecg
-	Max_r

-	ex_ang
-	ST_diff
-	Slope
-	Diab_num
-	Thalass

My approach was to build a full unnested model consisting of all variables followed by models based on the individual groupings and compare the different results. The 1st model, as shown in the table above, showed significance at the 5% level for employment, rest_ecg, ex_ang, ST_diff, Diab_num and Thalass.

The first step was to check for and deal with any missing data. Fortunately, the dataset was complete with no missing values so the data did not be altered. The lack of missing data is also why a complete case comparison analysis was not done, as there was no “missing” model to compare it to. After checking the presence of missing data, I decided to relevel all categorical variables to the respective level that had the most observations in each. The decision to also centre the continuous predictors was made to see if there was any meaningful change in their significance. Age, Resting BPS, Cholesterol and Max HR are all continuous variables that do not have any meaningful intercept interpretation, as, in each case, a figure of 0 would mean the patient is non-living. However, centring in the full model did not seem to help the data become any more information for these variables so the original variables were kept.

Based on the groupings table above, two themed models were built. In the demographic model, only Employment showed significance. This cemented employment as an important factor as in both the full and nested model, it stood out as a predictor. For the physical factors, Chest pain, St diff, slope, diab_num and thalass were all significant, suggesting that biological factors were more explanative than their demographic counterparts. Therefore, when looking at which variables to include in the later models, these were of particular interest. Neither of these nested models, by the deviance test or by using a classification table, were better than the full model.

The next few models that were built were based on my initial thoughts in the explanatory analysis after viewing the plots; the predictors were combinations of the variables that I thought would have some relationship to CVD. After doing that and removing non-significant predictors at each stage, I settled on a model consisting of employment, chest_pain, rest_ecg, ex_ang, ST_diff, diab_num and thalass as the final variables. From a deviance standpoint, this model was not better than the full model, however, the classification table showed that it predicted the observations with slightly better accuracy than the first model for this dataset. The first (full) model predicted 0s/1s with 93%/93% accuracy, while the new model predicted with 93/94% accuracy.

It was at this point that interactions were tested to see if a model could be built that passed both evaluation tests. The interactions created aliased coefficients, however, examining the gvif of the model with no interactions showed that the variables themselves were not collinear. The interactions increased the accuracy of prediction, peaking at 94/95% in my explorations.

Outliers were tested by examining the outlier statistics (leverage values, Diffts, standardised residuals and cook’s distance) and looking for any intersections of all four, to which there were none. Therefore, no data point was removed due to it being an anomaly.

words:533

Final Results

```
Final.model<-glm(cvd~employment+chest_pain+rest_ecg+ex_ang+ST_diff+diab_num+thalass+chest_pain*ex_ang+employment*chest_pain,data= data2, family=binomial(link="logit"))
```

	Final Model	P-values
Intercept	4.87	0.000201
employment2	-0.007385	0.954927
employment3	-4.87	0.000379
employment4	-6.974	0.0000409
chest_pain1	-1.116	0.511285
chest_pain2	18.45	0.993180
chest_pain3	35.24	0.995469
rest_ecg1	2.492	0.001980
rest_ecg2	4.631	0.182382
ex_ang1	-2.837	0.006211
ST_diff	-8.034e-01	0.011593
diab_num1	-2.275	0.002044
diab_num2	-3.448	0.001913
diab_num3	-5.013	0.011961
diab_num4	1.41	0.897026
thalass1	-2.208	0.136392
thalass3	-1.840	0.005840
chest_pain1::ex_ang1	5.351	0.027675

chest_pain2::ex_ang1	-2.089	0.955269
chest_pain3::ex_ang1	-1.599	0.997192
employment2::ches_tpain1	-2.998	0.153181
employment3::chest_pain1	2.658	0.236711
employment4::chest_pain1	3.364	0.146780
employment2::chest_pain2	-17.79	0.993423
employment3::chest_pain2	-19.25	0.992884
employment3::chest_pain2	-14.05	0.994807
employment2::chest_pain3	NA	NA
employment3::chest_pain3	-31.84	0.995905
employment4::chest_pain3	-53.21	0.996388

The analysis settled on the above model as the final model. This model provides quite a good fit for the data, as its classification table below shows. The diagnostics were also good, as the deviance test is successful, the residual deviance is significantly different from the null deviance, showing that this model is preferable to its null model.

	obs=0	obs=1
pred=0	130	9
pred=1	8	156
%correct	94	95

Considering employment, being high skilled versus all other lower-skilled levels of employment results in an odds ratio of 9.36e^-4, meaning that those who are highly skilled are much less likely than lower-skilled workers to get CVD. When looking at chest_pain, we see that those who have no symptoms or chest pain unrelated to typical angina, patients are much more likely to have CVD - the two have odds ratios of 102975329.7 and 2.01622E+15 respectively. Comparing this to atypical angina, which has an odds ratio of 0.3275875275, this suggests that angina is not typically indicative of CVD as a whole,

Considering the number of fluorescent-coloured blood vessels, we see the odds of having CVD reduce as the number of fluorescently coloured blood vessels increase up until the fourth level, where the odds ratio of having CVD spike to 4.095955404. This suggests that having these types of blood vessels isn’t typically dangerous until there are too many of them, leading to a 4 times higher chance of having CVD. Looking at blood thalassemia shows an interesting connection. Those without a fixed defect are generally less likely to have CVD. Those with no defect are about 90% less likely to have CVD, while those with a reversible defect are about 85% less likely to have CVD. This shows that blood thalassemia is a serious defect that can lead to worse diseases, especially if the defect is serious.

We see the same type of relationship with rest_ecg. Compared to having normal ECG readings, those with abnormal ST-T or possible left ventricle hypertrophy are much more likely to have CVD. Those with abnormal ST-T are about 12 times more likely while those with possible ventricle hypertrophy are 102 times more likely to have CVD. When looking at the ST difference closely, we see that on average, a 1 unit increase in the ST difference leads to a 66% decrease in the odds of having CVD, which matches my prediction in the explanatory analysis that those with a shorter ST_diff are more likely to have CVD.

Interestingly, having angina in the past actually decreases the chances of CVD by 95% compared to those who did not in this example.

Words:369

Lay Report:

How to know if you are at risk of Cardiovascular disease?

One point that medical officials have become increasingly aware of is the prevalence of lifestyle diseases, specifically those related to diet choices such as cardiovascular disease. Many are advised to watch their weight, exercise more and reduce bad lifestyle habits in order to reduce the chance of suffering from this disease. However, what truly affects CVD, and how can you know if you are at risk? You might be surprised to know that even your job could affect how likely you are of suffering from the disease. So what truly matters when talking about CVD?

What physical factors should I be wary of?

A study was done on several patients (both male and female ranging between the ages of 29 and 77) who had reported chest pain/discomfort to their GPs, a common symptom of CVD, who were then referred to a consultant to run further biological tests. Of these patients, those who also had blood thalassemia, a disorder that causes the body to produce less haemoglobin, and those that reported a fixed defect were much more likely to have CVD. In fact, those who had only reversible defects or no defect at all were 85% and 95% less likely to have CVD than those who did have a fixed defect. It might be worth checking if you have this condition at your next check-up if you feel that you may be at risk or have had consistent chest pains recently. Other biological factors also seemed to show similar relationships. For example, those who had 4 or more fluorescently coloured blood vessels were about 4 times as likely to have CVD than those with none of such blood vessels. You might want to watch out for these too. Resting ECG categorisations of the patients showed that those that showed atypical ST-T readings on the ECG (ST and T are phases of the heartbeat) were about 12 times as likely as no defect patients to have CVD, while those that showed signs of left ventricle hypertrophy, a severe condition, were over 100 times as likely to have CVD.

We can look at the heartbeat more closely through the ST difference data, which showed that having a shorter distance between these two phases leads to an increased chance of CVD. In fact, each unit increases the vertical distance between the phases, leading to a 65% decrease in the chance of having CVD than previously. Shorter ST differences are associated with several heart defects, so if you fall into this category it might be indicative of a more severe, underlying heart disease. Also, for those that are experiencing chest pains, the study found that if these chest pains are unrelated to angina, they result in severely higher chances of CVD. This coincides with data focused on those that had angina in the past, as those who have had angina were reported to be at a lower risk than those that did, with approximately 95% less of a chance.

The study also found a relationship between the level of employment and CVD. It seemed that those who work in lower-skilled professions, ranging from manual labour to mid-skilled services, were much worse off compared to their higher-skilled counterparts. Mid-skilled patients, for example, low-level managers in firms, were 99% less likely to have CVD than manual labourers. The chance was almost 100% for high skilled professions. Other low-skilled professions such as entry-level positions were closer to the lowest level, however, even these were still about 1% less likely to have CVD than unskilled labourers. This data shows that the skill level of the profession, usually associated with career progression and financial capability, has a noticeable effect on whether a patient is at risk of CVD or not. This does not exempt higher-skilled from the disease, however, it is much more prevalent in lower-skilled employees.

Can I trust this data?

Overall, the sample, being only just over 300 participants, is quite small and, therefore, we must ask how useful these results are, as they may not be representative of the whole population. In this dataset, the patients had already been referred after having complained of some level of chest discomfort. Many people who have CVD do not have any symptoms and, therefore, do not know that they have heart problems. If the results of the study were abstracted to the entire population, it is possible that many with such risks might be missed due to a lack of symptoms such as chest pain, which is a limit of the study itself. Overall, because the sample size is small, it is possible that with more patients, better parallels could have been drawn between the biological and demographic factors that can predict CVD. However, that is not to say that we should ignore these findings, as they highlight many useful relationships with factors that are related to and are affected by lifestyle habits. Therefore, in conclusion, if you notice consistent chest pains, it is advised to have a medical professional run a series of tests including the ones listed above as CVD might be a real risk.

Word count: 859

Appendix;

There was no missing data so I did not have to remove levels. I also did not have to merge any levels as no one variable had too many levels or levels that I thought should be merged together,

Variable	Continuous or Categorical	Changes +reasons

Employment	Categorical	Relevelled to 1 (unskilled workers) to compare higher skill levels versus the lowest level
Chest_pain	Categorical	Relevelled to 0 (typical angina) to compare angina to other types of chest pain
Rest_ecg	Categorical	Relevelled to 0 (normal) in order to compare normality and abnormalities.
Ex_ang	Categorical	Relevelled to 0 (no)
ST_diff	Continuous	
Diab_num	Categorical	Relevelled to 0 to compare having no fluorescently coloured blood vessels versus actually having them
Thalass	Categorical	Relevelled to 1(normal) for abnormality investigations