Bayesian Networks: Assignment 1

Ron Hommelsheim
s1000522

Marlous Nijman
s4551400

Steffen Ricklin
s1009136

December 1, 2020

Contents

Abstract

Heart disease is the leading cause of death globally. Different symptoms and clinical measurements can be used to predict whether a patient has heart disease. In our report, we tried to model the underlying causes that influence a persons chances of heart disease. Results show that our model predicts heart disease in patients with about 80% accuracy.

1 Introduction

Cardiovascular diseases (CVDs) are the primary cause of death globally. An estimated 17.9 million people die each year, according to the World Health Organization (WHO) [1]. This represents 31% of all global deaths, 85% of which are due to heart attack and stroke. Investigating the symptoms which indicate those who are at the highest-risk, can ensure early treatment and possibly prevent death. Therefore, it is paramount to research the different conditions that explain heart disease. In this research we used a data-set of 14 attributes [2] to create a causal diagram. The initial structure was created by consulting the literature on the specific attributes, to understand probable causal relationships. Continuous data was transformed to categorical in order to test the network structure by applying Chi-square tests. As a third step, we pruned the Bayesian network and compared how well the three distinct network structures explain the data and predict the diagnosis. We conclude with a discussion of our methods and suggest strategies for improvement.

2 Methods

First, we created an initial causal diagram. We tested its structure using the Chi-square test. Accordingly, we added edges to accommodate the test's implications. Lastly, we determined edge coefficients and pruned superfluous edges. The data was fitted on all three network structures to determine plausibility. The detailed methodology is described below. All implementations were done in the programming language ”R”. We made use of several libraries, namely ”dagitty”, ”bayesianNetworks”, and ”bnlearn” for the network structure and functions on the network such as the testing of implication dependencies, ”pROC” to analyse results, ”ggplot2” for plotting and ”caret” for k-fold cross validation and for the confusion matrices.

2.1 Data

To investigate the potential causes and symptoms that best predict heart disease, the Heart Disease Data-Set from the UC Irvine Machine Learning Repository was used [2]. The original data-set contains a total of 76 attributes, but we used a subset of 14 variables that all published experiments refer to [2]. More specifically, we used the processed Cleveland data-set which already contains only the subset. The variables we used are summarized in table 1 in the appendix. First we expound on some of the perhaps less known attributes as to get a clear understanding of what they entail before delving into the construction of the network.

2.1.1 Thalassemia

Thalassemia is an inherited blood condition. It is autosomal recessive which renders it independent of gender. A person with Thalassemia has fewer red blood cells and less hemoglobin than the body should have. Hemoglobin lets red blood cells carry oxygen to all parts of the body. This can cause anemia, which is a condition with that has the feeling of tiredness as a symptom [3, 4].

2.1.2 Cholesterol

Cholesterol is a lipid. An excess of cholesterol in the blood, leads to clogging of the arteries, causing a process called atherosclerosis, a form of heart disease. The arteries become narrowed and blood flow to the heart muscle is slowed down or blocked. The blood carries oxygen to the heart, and if not enough blood and oxygen reach the heart, it can cause severe chest pain. If the blood supply to a portion of the heart is completely cut off by a blockage, the result is a heart attack [5, 6].

2.2 Pre-processing

Out of the 14 variables, 5 were continuous which we binned according to their respective ranges and distributions, such that the bins are balanced. Besides, we tried to keep number of bins small to so that the degrees of freedom during testing our network will not be too high. Furthermore, we also binned the categorical variable 'diagnosis' into two bins instead of 5. This was done because the label 0 that indicates absence of heart disease occurred much more often compared to the labels 1 to 4 that indicate presence of heart disease. The labels 1 to 4 were therefore binned in a single bin. Next to binning, there were also some missing values in the data. We dealt with this by assigning those the value of the most occurring value for the given variable. Since the data-set was already processed, we only needed to bin variables to have them in in a format such that we can test our network structure. See Table 1 to see the binned variables.

2.3 Initial Construction of the Causal Diagram

In order to accommodate causal relationships in the Bayesian network, attributes were connected according to the time of their measurement and by consulting the literature. Sex, age, and Thalassemia are determined at birth and are therefore root nodes. Chest pain and exercise induced angina are both symptoms or conditions which a person can notice in day-to-day life. The resting ECG, as well as the maximum heart rate, resting blood pressure, cholesterol level, and fasting blood sugar are all measured at the doctor. The ST-depression, the ST-slope, and the coloured arteries are all consequences of the measurements and therefore happen after them. Lastly, there is a diagnosis.

sex → cholesterol: Before menopause, women tend to have lower total cholesterol levels than men of the same age. After menopause, however, women's cholesterol levels tend to rise [7].

sex → max heart rate, resting blood pressure, fasting blood sugar, chest pain: There is evidence that the maximum heart rate as well as the resting blood pressure, fasting blood sugar, and chest pain depend on gender [8, 9].

sex → cholesterol: Sex as a predictor of cholesterol levels have been documented by Schaefer et al. [7].

age → resting ecg, rest. blood p., chest pain, fast. blood sugar, max heart rate: Research done by Tanaka et al. [10] shows that the maximum heart rate is strongly related to age while research by Landahl et al. [11] shows how blood related variables change with age. Plasma glucose levels progressively increase with age in Hong Kong Chinese non-diabetic subjects [12].

age → exercise induced angina: Angina is typically rare in persons under the age of 35 [13].

age → cholesterol: As we get older, cholesterol levels rise according to a meta-analysis done in 2007 [14].

thalassemia → exercise induced angina: Thalassemia is a genetic blood disorder that impacts the ability of the blood to get oxygen to the body's organs. Thalassemia is not age or gender related. But can cause angina due to the low amount of red blood cells.

thalassemia → max heart rate, rest blood press: Symptoms common to many types of anemia include easy fatigue, loss of energy, unusually rapid heart beat, particularly with exercise [3].

chest pain → rest ecg, max heart rate, resting blood pressure: Thompson et al. [15] show that people with chest pain is an indicator of variance in the blood flow.

chest pain → cholesterol: High cholesterol levels often have chest pain as a consequence [16].

chest pain → diagnosis: Chest pain is also a robust indicator of heart disease.

exercise induced angina → rest ecg, max heart rate, rest blood pressure: Multiple studies (e.g. [17])show how angina affects blood flow related variables.

exercise induced angina → cholesterol: Most people with angina have either elevated blood pressure or cholesterol or a combination of both [6].

resting ecg → ST-depression, ST-slope: ST depression and slope are the curve the ecg displays.

max heart rate → ST-depression, ST-slope: ST is the result of the heart rate measurement.

max heart rate → diagnosis : The maximal heart rate has been shown to be a decent indicator of congenital heart disease [18].

resting blood pressure → ST-depression, ST-slope: The resting blood pressure clearly affects the measurement of the heart rate.

resting blood pressure → coloured arteries: Colour Doppler ultrasound is used to examine the velocity of blood flow.

resting blood pressure → diagnosis: MacMahon et al. show that the resting blood pressure is a predictor of coronary heart disease [19].

cholesterol → coloured arteries: The colouration can show whether cholesterol has blocked any arteries.

cholesterol → diagnosis: Cholesterol is blocking arteries which is a direct cause of heart disease [20].

Fasting blood sugar → diagnosis: Fasting blood sugar levels are robust predictors of diabetes and of heart disease [21].

ST-slope, ST-depression → diagnosis: ST slope and depression analysis has been shown to improve the prediction of all-cause and cardiovascular mortality [22].

coloured arteries → diagnosis: Colouring the arteries is a technique, specifically for the diagnosis of heart disease.

See the full network structure in Figure 2 in the appendix.

2.4 Improving and testing the network

Using the initial network structure from figure 2 several methods were applied to test and improve this first network. The procedures described below result in three different networks.

Since all non-categorical variables were converted into categorical variables, the Chi-square test was the chosen method for testing the conditional independences within our network. To acquire the test results for all possible conditional independences we used the `localTests` function from the R package dagitty, e.g.:

```
localTests(net, data, type="cis.chisq")
```

2.4.1 Conditional independence tests

The p-values and the RMSEA values obtained from the chi-square localTests function were used to adjust network 1's structure. If the test for that conditional independence were significant (p-value < 0.05), then a RMSEA value > 0.04 indicated that adding an edge between the tested variables might improve the network. However, RMSEA values become more unreliable the higher the number of conditioning variables is. Therefore we started with a low number of maximal conditioning variables. This limit was increased, once all significant tests are dealt with. For cases in which a test indicated that two variables should be independent, although the literature offers clear arguments against the dependence relation, we addressed this conflict.

After applying the chi-square test to the initial network, we obtained an adjusted network (network 2, see figure 3) that is more interconnected and has fewer significant independence relationships than network 1. Using network 2, we predicted its performance by fitting the network and predicting its accuracy using k-fold cross validation. Furthermore, computing the correlation coefficients indicates which variables were the most influential predictors of heart disease in the network given the data.

2.4.2 Pruning

In further steps, the correlation coefficients are used to prune the adjusted network. By only keeping those edges for which the coefficients are significant (p-value ≤ 0.05), the resulting network (network 3, see figure 4) contains only the most influential edges.

2.4.3 Comparing the networks

Since the Cleveland data set that was used to train and test the three different networks contains only 303 observations, simply training once and computing the accuracy values might be imprecise. Therefore, the k-fold cross validation method was used to receive stable accuracy values. Furthermore, confusion matrices and ROC curves were constructed to better compare the networks' performances.

3 Results

First, we will show which edges were added or removed from the initial network to end up with our adjusted network. To analyse the different network structures that we built, we compared the performance of each network based on predictions made on a test set using k-fold cross validation, as well the coefficients of the network.

3.1 From Initial to Adjusted Network

As explained in the method section, we looked at the p-value as well as to the RMSEA of the (conditional) independencies to determine where to add edges. We will go through the results by increasing the number of max-conditioning variables. We will list the edges that were added, but we will not show all related p-values and RMSEA scores for simplicity. These results can be replicated by running our code, which can be found on GitHub [1].

max-conditioning variables = 1

- We first found that the independence between sex and thalassemia was unlikely according to the data . However, since this is in contrast with the literature [23], no edge was added between these variables
- sex → exercise induced angina

[1] https://github.com/R1704/heart-disease

max-conditioning variables = 2

- fasting blood sugar → ST-depression
- chest pain → coloured arteries
- fasting blood sugar → coloured arteries

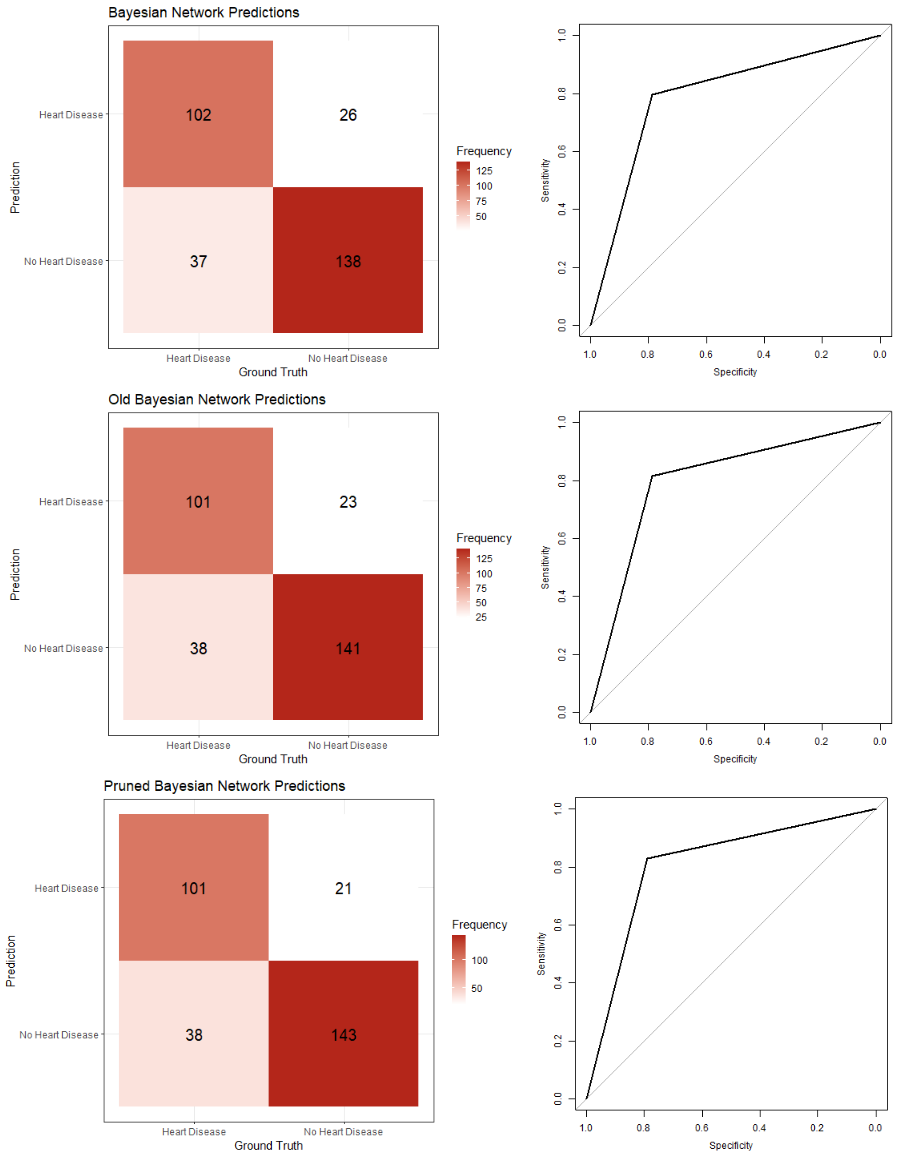

Figure 1: Confusion matrix and ROC curve for our initial network (top), adjusted network (middle), and pruned network (bottom).

max-conditioning variables = 3

- ST-slope → ST-depression
- exercise induced angina → chest pain
- thalassemia → ST-slope

max-conditioning variables = 4

- age → coloured arteries
- exercise induced angina → ST-slope

We stopped adding edges after a max-conditioning variables of 4, because the network was already quite dense, and the Chi-square test becomes unreliable when there is a large number of conditioning variables. We then moved on to fitting the model and observed that cholesterol has a very low effect on the diagnosis (-0.0048). Therefore, we removed the edge from cholesterol to diagnosis.

3.2 Predictions

First, the adjusted network achieved an accuracy of 0.792 and AUC of 0.79. The initial network has an AUC of 0.8 and an accuracy of 0.799. Finally, the pruned network has an accuracy of 0.799 and an AUC of 0.81. Figure 1 shows the confusion matrix and ROC curve for each of our tested networks.

3.3 Coefficients

Next to predictions, we also inspect the coefficients of the networks to compare them. The coefficients from all nodes that are connected to diagnosis can be found in Appendix C.

4 Discussion

Although we tried to create the initial network as accurately as possible by reading the literature thoroughly, we are by no means experts and therefore the proposed causal relationships could be faulty. To improve it, it would be necessary to consult experts on coronary disease. The sequential strategy of building a network structure, testing the network and adjusting it and lastly pruning it gave us an insight and was advantageous. Sometimes counter-intuitive connections in the network seem to explain the data rather well. Pruning the network follows the Occam's razor principle: If it is possible to explain the same thing in simpler terms it is advised to use the simpler model.

The accuracy and AUC of all three models are very similar, even though the models are all quite different. Overall, the pruned network seems to perform slightly better compared to the other two models, with an AUC of 0.81.

In general, our models all seem to perform fairly well, with an accuracy around 80%. We say this, because our data-set was very small: it only contains 303 data points. The performance is not good enough for the model to be used in real clinical settings. However, it does seem

to explain the data fairly well.
The coefficients that we computed gave us some insight into which variables are important predictors for heart disease. The most important variables across all networks seem to be chest pain and the number of coloured arteries (with coefficients 0.29 and 0.36 in the pruned network, respectively). The number of coloured arteries are the number of major vessels colored by fluoroscopy. Furthermore, we see that removing the edge from cholesterol was a good choice, since its coefficient is only -0.01. Finally, we see that the two edges with the lowest coefficients in the adjusted network are removed in the pruned network (fasting blood sugar (-0.05) and rest blood pressure (0.10)).

5 Conclusion

In this project we modeled the underlying causes of heart disease according to the literature and our own insights, based on testing of our network. Our networks all achieved an accuracy of around 80%, which is relatively good given our limited amount of data.
Despite the networks showing similar performances, one might argue that the simpler network is the best choice here, which would be the pruned network, which only takes into account the most probable edges.

References

[1] World Health Organization. Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[2] UC Irvine Machine Learning Repository. Heart disease data set. http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[3] Kinderblutkrankheiten. Symptome bei thalassaemia major. (https://www.kinderblutkrankheiten.de/content/erkrankungen/rote_blutzellen/anaemien_blutarmut/thalassaemie/symptome/thalassaemia_major/#:~:text=Die%20Milz%20vergr%C3%B6%C3%9Fert%20sich%20dabei,kommen%20kann%20(pulmonale%20Hypertension).

[4] Centers for Disease Control and Prevention. What is thalassemia? https://www.cdc.gov/ncbddd/thalassemia/facts.html.

[5] National Center for Biotechnology Information. Cholesterol. https://pubchem.ncbi.nlm.nih.gov/compound/Cholesterol.

[6] John M Chapman, Anne H Coulson, Virginia A Clark, and E Raymond Borun. The differential effect of serum cholesterol, blood pressure and weight on the incidence of myocardial infarction and angina pectoris. *Journal of chronic diseases*, 23(9):631–645, 1971.

[7] Ernst J Schaefer, Stefania Lamon-Fava, Susan D Cohn, Mary M Schaefer, JM Ordovas, WP Castelli, and PW Wilson. Effects of age, gender, and menopausal status on plasma

low density lipoprotein cholesterol and apolipoprotein b levels in the framingham offspring study. *Journal of lipid research*, 35(5):779–792, 1994.

[8] Marcus W Agelink, Rolf Malessa, Bruno Baumann, Thomas Majewski, Frank Akila, Thomas Zeit, and Dan Ziegler. Standardized tests of heart rate variability: normal ranges obtained from 309 healthy humans, and effects of age, gender, and heart rate. *Clinical Autonomic Research*, 11(2):99–108, 2001.

[9] Dan Ziegler, G Laux, K Dannehl, M Spüler, H Mühlen, P Mayer, and FA Gries. Assessment of cardiovascular autonomic function: age-related normal ranges and reproducibility of spectral analysis, vector analysis, and standard tests of heart rate variation and blood pressure responses. *Diabetic Medicine*, 9(2):166–175, 1992.

[10] Hirofumi Tanaka, Kevin D Monahan, and Douglas R Seals. Age-predicted maximal heart rate revisited. *Journal of the american college of cardiology*, 37(1):153–156, 2001.

[11] STEN Landahl, Calle Bengtsson, JOHAN A Sigurdsson, ALVAR Svanborg, and K Svärdsudd. Age-related changes in blood pressure. *Hypertension*, 8(11):1044–1049, 1986.

[12] Joyce SF Tang Gary TC Ko, Hendena PS Wai. Effects of age on plasma glucose levels in non-diabetic hong kong chinese. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2080461/.

[13] MD David Laxson. Five things to know about angina. https://www.mhealth.org/blog/2017/february-2017/five-things-to-know-about-angina#:~:text=Angina%20is%20rare%20in%20people,hypertension%2C%20smoking%20or%20high%20cholesterol.

[14] Prospective Studies Collaboration et al. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths. *The Lancet*, 370(9602):1829–1839, 2007.

[15] David R Thompson, Roger L Blandford, Terence W Sutton, and Paul R Marchant. Time of onset of chest pain in acute myocardial infarction. *International journal of cardiology*, 7(2):139–146, 1985.

[16] David W Scott, Antonio M Gotto, James S Cole, and G Anthony Gorry. Plasma lipids as collateral risk factors in coronary artery disease—a study of 371 males with chest pain. *Journal of chronic diseases*, 31(5):337–345, 1978.

[17] Lucien Campeau. Grading of angina pectoris. *Circulation*, 54(3):522–523, 1976.

[18] Gerhard-Paul Diller, Konstantinos Dimopoulos, Darlington Okonko, Anselm Uebing, Craig S Broberg, Sonya Babu-Narayan, Stephanie Bayne, Philip A Poole-Wilson, Richard Sutton, Darrel P Francis, et al. Heart rate response during exercise predicts survival in adults with congenital heart disease. *Journal of the American College of Cardiology*, 48(6):1250–1256, 2006.

[19] Stephen MacMahon, Richard Peto, Rory Collins, Jon Godwin, J Cutler, Paul Sorlie, Robert Abbott, J Neaton, Alan Dyer, and Jeremiah Stamler. Blood pressure, stroke, and coronary heart disease: part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *The Lancet*, 335(8692):765–774, 1990.

[20] Scott M Grundy. Cholesterol and coronary heart disease: a new era. *Jama*, 256(20):2849–2858, 1986.

[21] OP Agarwal. Prevention of atheromatous heart disease. *Angiology*, 36(8):485–492, 1985.

[22] Peter M Okin, Richard B Devereux, Jan A Kors, Gerard van Herpen, Richard S Crow, Richard R Fabsitz, and Barbara V Howard. Computerized st depression analysis improves prediction of all-cause and cardiovascular mortality: the strong heart study. *Annals of noninvasive electrocardiology*, 6(2):107–116, 2001.

[23] Antonio Cao, Luisella Saba, Renzo Galanello, and Maria Cristina Rosatelli. Molecular diagnosis and carrier screening for β thalassemia. *Jama*, 278(15):1273–1277, 1997.

Appendices

A Data

Variable Name	Description	Variable Type	Levels	New levels
age	age in years	continuous	[29, 77]	....
sex	sex	categorical (binary)	1 = male 0 = female	-
chest_pain	chest pain type	categorical	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic	-
rest_blood_press	resting blood pressure (in mm Hg)	continuous	[94, 200]	1 = [90, 120) 2 = [120, 140) 3 = [140, 200]
cholesterol	serum cholestoral in mg/dl	continuous	[126, 564]	1 = [100, 200) 2 = [200, 300) 3 = [300, 600]
fasting_blood_sugar	fasting blood sugar > 120 mg/dl	categorical (binary)	0 = false 1 = true	-
rest_ecg	resting electrocardiographic results	categorical	0 = normal 1 = having ST-T wave abnormality 2 = showing probable or definite left ventricular hypertrophy	-
max_heart_rate	maximum heart rate achieved	continuous	[71, 202]	1 = [50, 110) 2 = [110, 140) 3 = [140, 175] 4 = [175, 210]
exercise_induced_angina	exercise induced angina	categorical (binary)	0 = no 1 = yes	-
ST_depression	ST depression induced by exercise relative to rest	continuous	[0.0, 6.2]	0 = 0.0 1 = (0, 2.0) 2 = [2.0, 6.5]
ST_slope	slope of the peak exercise ST segment	categorical	1 = upsloping 2 = flat 3 = downsloping	-
coloured_arteries	number of major vessels (0-3) colored by flourosopy	ordinal	0, 1, 2, 3	-
thalassemia	thalassemia	categorical	3 = normal 6 = fixed defect 7 = reversable defect	-
diagnosis	diagnosis of heart disease	categorical	0 = absence 1, 2, 3, 4 = presence	0 = absence 1 = presence

Table 1: Subset of attributes from the Heart Disease data-set which are used in all published experiments. The Cleveland data set, as a subset of the Heart Disease data set, does only include the listed attributes within this table. The column levels shows the categories and value ranges of the Cleveland data set.

B Networks

B.1 Initial Causal Diagram

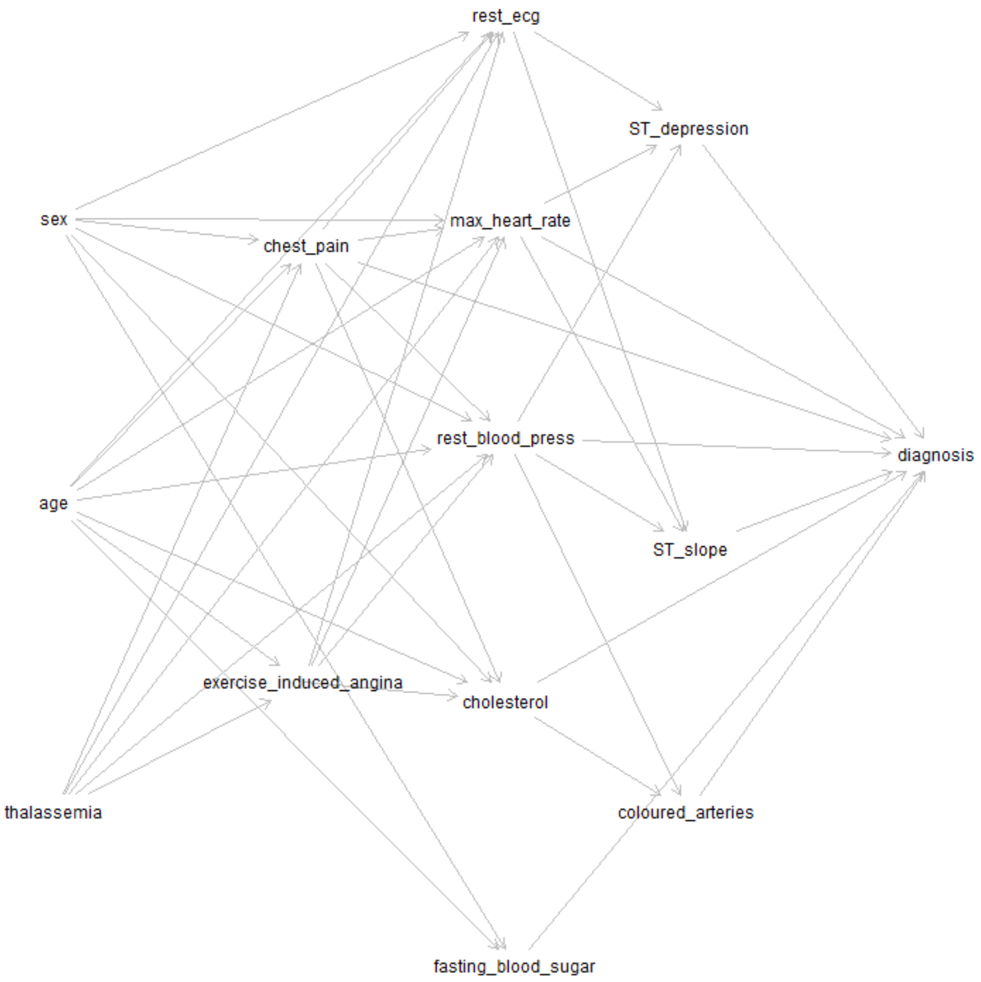

Figure 2: Initial network structure

B.2 Adjusted Network Structure

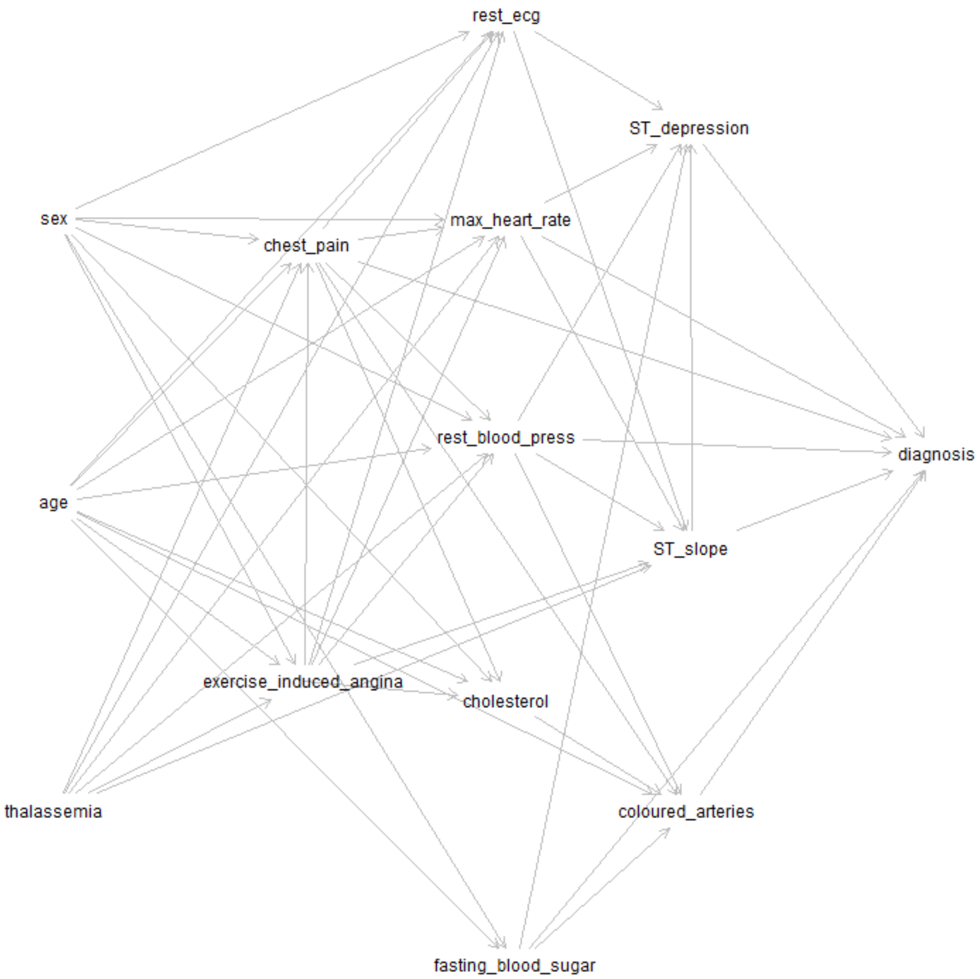

Figure 3: Adjusted Network Structure

B.3 Pruned Network

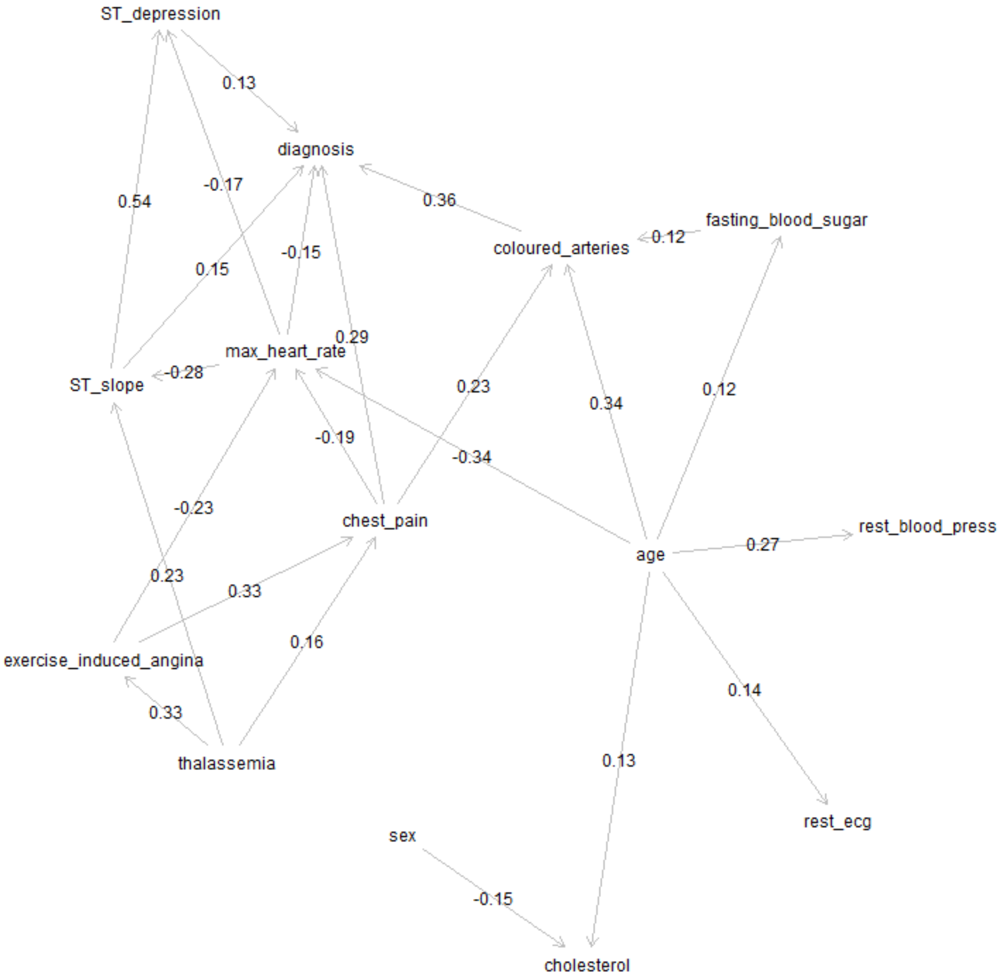

Figure 4: The Pruned Network

C Coefficients

C.1 Edge Coefficients

ST-depression	→ diagnosis	0.12
ST-slope	→ diagnosis	0.15
chest pain	→ diagnosis	0.30
cholesterol	→ diagnosis	-0.01
coloured arteries	→ diagnosis	0.35
fasting blood sugar	→ diagnosis	-0.05
max heart rate	→ diagnosis	-0.16
rest blood press	→ diagnosis	0.10

Table 2: Edge coefficients in the initial network

ST-depression	→ diagnosis	0.12
ST-slope	→ diagnosis	0.15
chest pain	→ diagnosis	0.30
coloured arteries	→ diagnosis	0.35
fasting blood sugar	→ diagnosis	-0.05
max heart rate	→ diagnosis	-0.16
rest blood press	→ diagnosis	0.10

Table 3: Edge coefficients in the adjusted network

ST-depression	→ diagnosis	0.13
ST-slope	→ diagnosis	0.15
chest pain	→ diagnosis	0.29
coloured arteries	→ diagnosis	0.36
max heart rate	→ diagnosis	-0.15

Table 4: Edge coefficients in the pruned network