Estimating the Causal Effect of Serum Cholesterol on Angiographic Coronary Disease

Bayesian Networks & Causal Inference - Report

Grønlund, Stian (s1122151) Jansen, Sjoerd (s1028353) Schöyen, Inga (s1127295)

Friday 29th November, 2024

1 Introduction

According to the American Center of Disease control (CDC), heart disease is the leading cause of death, causing about 1 in every 5 deaths in the US [1]. Similarly, in the EU 32.4% of all recorded deaths in 2021 were the result of cardiovascular disease, which represents the main cause of death in all member states [2]. Despite the prevalence of heart disease, it is not clear what it is caused by and how strong those effects are. One of the main risk factors, at least according to public discourse, is the level of cholesterol in one's blood, i.e. serum cholesterol [3]. There are obvious public and scientific pitfalls to believing such statement without the necessary statistical justification. A bias could come to exist in our conceptualisation of the risk factors and approaches for treatments, and lead to unnecessary interventions in the patients' metabolism. Additionally, factors like age and genetic predispositions are commonly found to contribute to the risk of developing heart disease [3], and likely interact with the effect of cholesterol on this risk. Therefore, studying the causal structure of the variables that are associated with developing heart disease, and especially cholesterol, could make disease prevention more effective and cost and time efficient.

1.1 Aim of the Project

As serum cholesterol is a classic risk indicator of heart disease [3], we want to ascertain the causal effect of high cholesterol levels on heart disease risk. To do so, we will construct a causal, graphical model of the relationships between variables, and use this model to estimate the causal effects of certain predictor variables on the risk of heart disease. We focus on a focal relationship between cholesterol and heart disease, but also that of other ancestors of heart disease in the graph, as the relative contribution between these (and cholesterol), is important for future decisions about which interventions to prioritize. In particular the relation with fasting blood sugar sparks interest. We hypothesize that high fasting blood sugar is a cause of heart disease and is caused by high cholesterol levels. In short, we pose the following research questions:

- To what extent does high cholesterol lead to heart disease?
- What are the relative contributions to heart disease of age, sex, cholesterol, and fasting blood sugar?

2 Methods

We use the Heart Disease dataset [4], a UCIML dataset. This dataset has 303 instances of 68 variables of which only 14 are commonly used (listed in Table 3). Of these 14 variables, 13 are "predictor" variables, and the "outcome" variable diagnosis of heart disease. Table 3 shows all variables, with their type of data, range, and description, as well as the acronym we chose to represent it as in the DAGs.

The code used for the analysis and creation of the plots is accessible on github. A `Python` script utilizing the `ucimlrepo` package was used to fetch the dataset and save it in `CSV` format. The analysis was done using R scripts and the `lavaan[5], dagitty[6]` and `ggdag[7]` libraries (specific functions that were used will be mentioned in the corresponding sections).

In the following, we outline the steps we took to pre-process the data for the analysis, to construct the starting DAG and adjust it based on the assumption tests, and the final evaluation of the SEM based on the DAG.

Variable Name	Type	Range	Description	Acronym
subject_id	Integer	0-303	ID of the subject	ID
age	Integer	29-77	Age of the Subject	AGE
sex	Categorical	[0,1]	Sex of the Subject	SEX
cp	Categorical	[1,2,3,4]	Chest Pain Type	CP
trestbps	Integer	94-200	Resting blood pressure	BPr
chol	Integer	126-564	Serum cholesterol	Chol
fbs	Categorical	[0,1]	Fasting blood sugar	FBS
restecg	Categorical	[0,1]	Categorical Outcome of Resting ECG	ECGr
thalach	Integer	71-202	Maximum heart rate achieved during the Thalach Test	HRmax
exang	Categorical	[0,1]	Exercise induced angina	ANGe
oldpeak	Float	0.0-6.2	ST depression induced by exercise relative to rest	STd
slope	Categorical	[1,2,3]	Slope of the ST curve	STs
ca	Integer	[0,1,2,3]	Number of blood vessels counted	CA
thal	Categorical	[3,6,7]	Categorical Outcome of the Thalach Test	Thal
num	Categorical	[0,1]	Diagnosis of Heart Disease	HD

Table 1: Details about the Dataset Variables

2.1 Data preprocessing

The data required some pre-processing before we could perform the analysis with it. Of the 14 variables that we use, six were numeric, and eight categorical. We standardised all numerical data, that is, we scaled and translated all numeric data to have mean 0 and variance 1, to allow for scale-free comparison.

Of the categorical predictor variables, SEX, Fasting Blood Sugar (FBS), and Exercise-induced angina (ANGe) are binary, and were therefore left untouched. The remaining variables, Chest Pain (CP), Resting ECG (ECGr), ST-curve slop (STs), Blood vessel colouring (CA), and Thalach test (Thal) needed to be processed such as to allow linear modeling. We conceived of two approaches, which we pursued in two separate analyses, presented below. Given the different encodings, the analyses followed the same model construction and estimation outlined below.

The outcome variable HD had four categories, not explicitly explained or documented in the original paper or dataset description. We interpreted this as increasing severity of diagnosis and, following the example of previous works that used the dataset, decided to turn this into a binary variable for this analysis, indicating either no diagnosis (0) or a diagnosis level one or greater (1).

2.1.1 Approach 1

Name	Type	Mean	Variance	Number of levels	Levels
AGE	numeric	0.000	1.000	-	
SEX	binary	0.677	0.219	2	0—1
CP	binary	0.923	0.072	2	0—1
BPr	numeric	0.000	1.000	-	
Chol	numeric	0.000	1.000	-	
FBS	binary	0.145	0.124	2	0—1
ECGr	ordered	NA	NA	3	0—1—2
HRmax	numeric	0.000	1.000	0	
ANGe	binary	0.327	0.221	2	0—1
STd	numeric	0.000	1.000	0	
STs	binary	0.532	0.250	2	0—1
CA	ordered	NA	NA	4	0—1—2—3
Thal	ordered	NA	NA	3	0—1—2
HD	binary	0.946	1.524	2	0—1

Table 2: Variable Details Approach 1, Post-Processing

The first option we went with was to adapt the data in such a way that all variables could be considered ordered. This would open the way to using polychoric correlations to make statements about causal effects. In general, these polychoric correlations are easier to interpret, at the cost of losing some information about our data, and therefore making

interpretation harder again. Ultimately, we decided the benefit outweighed the cost, since most of the categorical variables could be binarized in a logical way. Hence, we proceeded to process the data as follows.

Firstly, we decided to turn CP, and STs into binary variables. CP originally had four categories; typical angina, atypical angina, non-anginal pain, and asymptomatic. We decided to code typical pain as 0, and all other levels as 1. The variable STs originally had three categories; upsloping, downsloping, and flat. In some research on this topic, downsloping and flat ST segments are considered similar [8], which lead us to decide to code upsloping as 0 and downsloping and flat as 1.

Secondly, we turned ECGr, CA, and Thal into ordered variables as follows. Resting ECG was turned into a factor with three levels, the original three categories had a sense of severity, which was used to order them. CA was reported as "number of vessels", and therefore allowed to be turned directly into an ordered factor with 4 levels. Lastly, Thal was turned into a factor with three levels. The original data was coded as; 3 - no defect, 7 - reversible defect, and 6 - fixed defect. These levels correspond to certain abnormalities during a thallium scintigram. We decided to code this as 0 - no defect, 1 - reversible defect, and 2 - fixed defect, as it seems that "fixed defect" is defined to be a "chronic" or "non-reversible" defect, i.e. an effect that showed up before or persisted after the stress test.

2.1.2 Approach 2

The alternative analysis is motivated by the observation that some of the categorical variables could also be perceived as unordered, namely CP, STs, and Thal. Instead of turning them into ordered variables or setting a threshold for binary encoding, the second analysis employs dummy coding for the unordered variables. To this end, we used the `dummycols` function from the `fastDummies` library to turn every category of each unordered variable into a binary variable. We chose to remove the most frequent category and set it as base level, which in all variables was the asymptomatic/normal condition.

name	**type**	**mean**	**var**	Number of levels	Levels
AGE	numeric	0.000	1.000	-	
SEX	binary	-0.323	0.219	2	0—1
BPr	numeric	0.000	1.000	-	
Chol	numeric	0.000	1.000	-	
FBS	binary	-0.855	0.124	2	0—1
ECGr	ordered	NA	NA	3	0—1—2
HRmax	numeric	0.000	1.000	-	
ANGe	binary	-0.673	0.221	2	0—1
STd	numeric	0.000	1.000	-	
CA	ordered	NA	NA	4	0—1—2—3
HD	binary	-0.054	1.524	2	0—1
CP_1	binary	-0.923	0.072	2	0—1
CP_2	binary	-0.835	0.138	2	0—1
CP_3	binary	-0.721	0.202	2	0—1
Thal_6	binary	-0.939	0.057	2	0—1
Thal_7	binary	-0.613	0.238	2	0—1
STs_2	binary	-0.539	0.249	2	0—1
STs_3	binary	-0.929	0.066	2	0—1

Table 3: Variable Details Analysis 1, Post-Processing

2.2 DAG construction

To construct an initial DAG for this dataset, we used theoretical knowledge and our own ideas to come up with a simple idea for a DAG. The plan was to make a DAG that had heart disease as the main node, i.e. to let it be a parent of most other nodes. Theoretically, this corresponds to viewing heart disease as a underlying cause and many other variables as possible symptoms, or outcomes of tests for it. Of course, many of the symptoms can have direct effects on each other. These effects are reflected in the model by attaching the corresponding directed edges to these nodes. Additionally, some of the variables are from the same tests, such as ECGr, STs and STd, so we chose to add edges between the nodes of these variables, as they can be expected to be correlated.

2.3 Model Testing & Adjustment

We tested our network structure in two phases, to evaluate whether the structure needs to be adjusted. First, we computed a polychoric correlation matrix M for our dataset using the `lavCor` functionality of `lavaan`. Using this matrix, we next

used the `dagitty` `localTests` function to test the measured covariances against the implied conditional independencies of our network model. Of these, we planned to make a small amount of changes to our network to remove the conditional independence assumptions that were most violated by the data, as judged by their p-value and effect estimate, without overcomplicating the network too much. After making these changes, we proceeded to fit the new model and examined the coefficients of the fit. For any of the coefficients that had an absolute value smaller than 0.01, we removed the corresponding edge from the graph. Following this, we finalised the DAG and proceeded with the effect estimation.

2.4 Effect estimation

With our final models we performed the causal effect analysis to estimate the effect of the predictors that were set to be exposure, namely AGE, SEX, FBS, and Chol, on the outcome HD.

For each of the predictors, we found an adjustment set using the `adjustmentSets` function from the `dagitty` library, to identify which covariates should be included in the regression for a given predictor. Based on the identified adjustment sets, we performed a binomial regression analysis using the `glm` function from the `stats` library, using the `family='binomial'` setting.

3 Results

We present the results of our project in the following section as follows. First, we describe the characteristics of the dataset, with respect to the distribution of variables and the correlations between them. Second, we describe the process of model testing and adjustment, and the resulting final DAG structures. And lastly, we describe the effect estimation of different predictor variables on heart disease (HD), both binary and ordinal encoded, based on the DAGs.

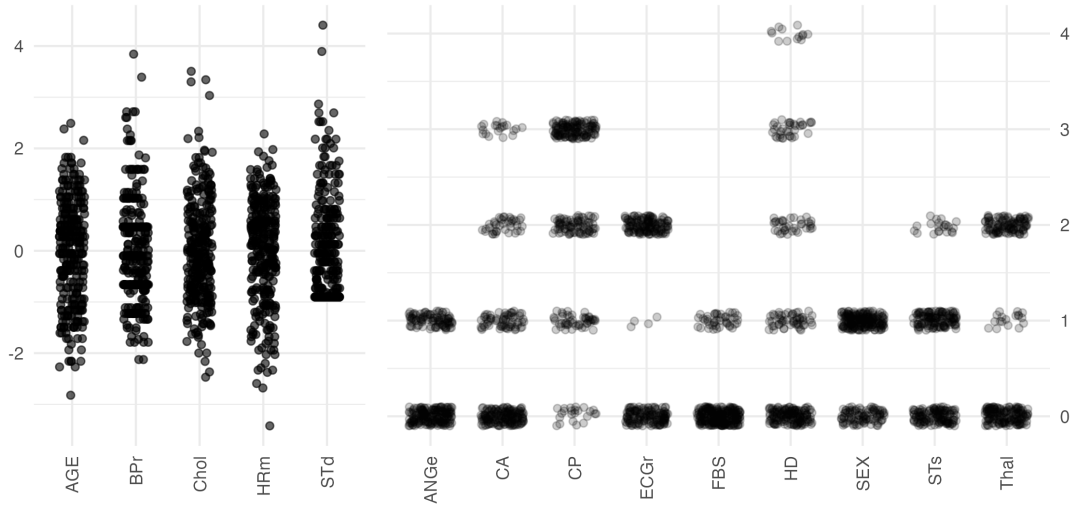


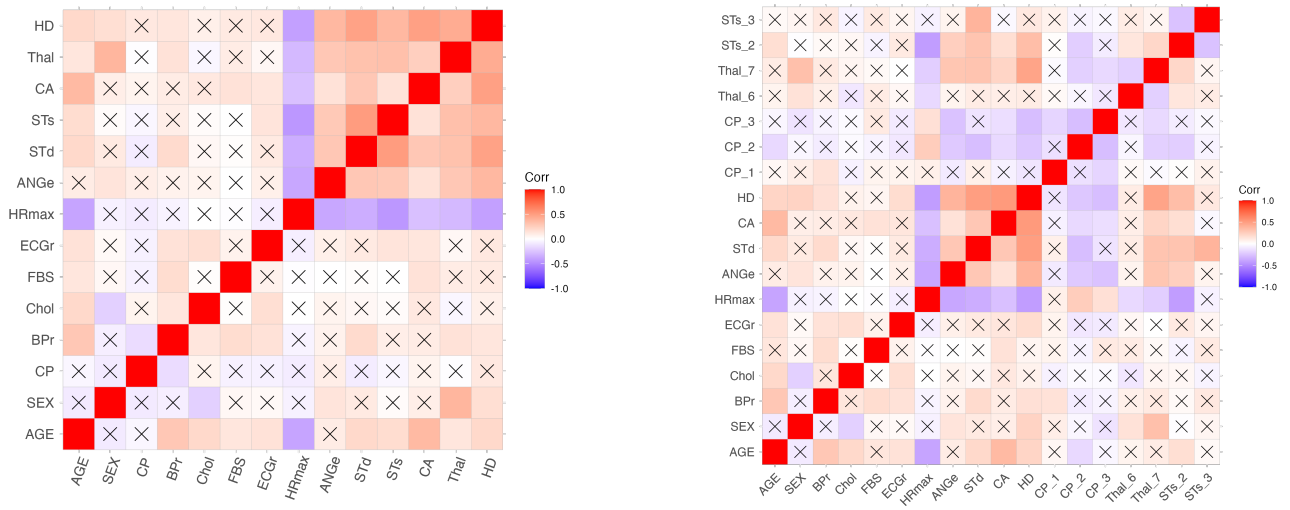
Figure 1: Scatter plots of all variables in the dataset

3.1 The Data

To get an understanding of the distribution of our variables and to check for outliers, we plotted all numeric variables in their standardised form, and all categorical variables starting from zero with one integer increments between the categories (see Figure 1). As can be seen, the continuous variables all seem roughly normally distributed, with the exception of STd, which due to its inherent cut-off at 0 is missing one tail. The categorical variables are more diverse, which some being more evenly distributed across the levels (ANGe, SEX), while others seem skewed to either end (CA, CP, FBS, HD, STs) or biased towards the extremes (ECGr, Thal).

Additionally, we looked at the level of correlation between all variables, with both encoding approaches (see Figure 2) using the `lavCor` function from the `lavaan` library. Some noteable trends include the significant positive correlation of AGE with most variables, except for HRmax, with which it is strongly negatively correlated. This connects to another trend of HRmax being exclusively negatively correlated with all variables. The outcome HD is significantly correlated with most variables, except CP, Chol, FBS, and ECGr, and with the exception of HRmax is exclusively positively correlated. A last noteworthy observation is that the dummy encoding reveals that some of the dummy coded variables that have

significant correlations as binary variables, have non-significant correlations for some of the levels while others remain significant.

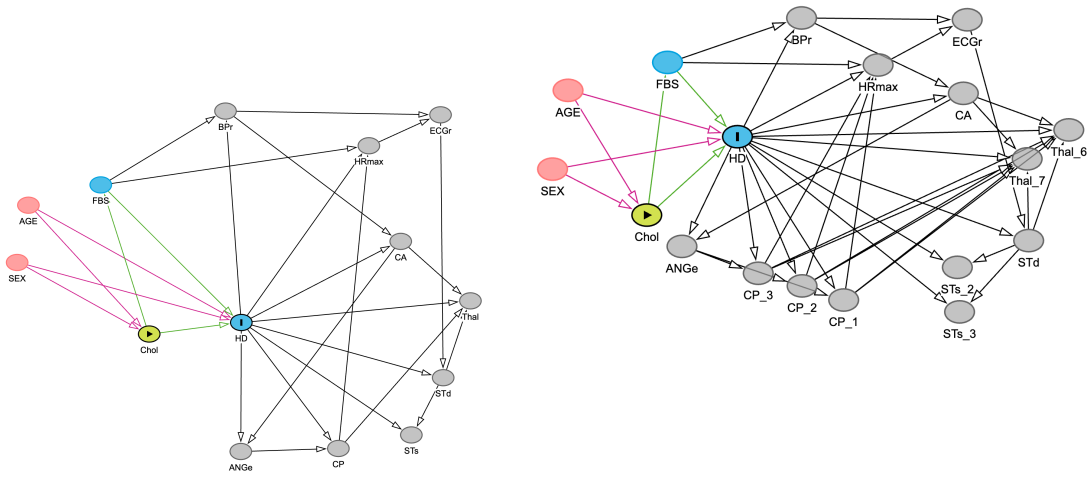


(a) Correlations with Approach 1 and HD binary

(b) Correlations with Approach 2 and HD binary

Figure 2: Correlation Matrix of Correlations between all variables, for both encoding approaches. Crosses signify non-significant correlation.

Both analyses started with the same DAG structure, depicted in Figure 4, which assumes that AGE, SEX, Chol, and FBS are ancestors of HD. All other variables are ancestors of HD, as they are outcomes of emergent characteristics of cardiovascular function and/or diagnostic tools for cardiovascular function. Amongst these, we connected the variables that are part of the same diagnostic procedure, like ECGr, STs, and STd, or seem to be logically connected, like ANGe and CP. To convert the DAG to include the dummy coding, that is to turn nodes representing variables into nodes representing levels, all edges of a variable node are preserved for all dummy levels, inheriting all ancestors and children.



(a) Approach 1

(b) Approach 2

Figure 3: Initial DAGs for both approaches

3.2 Analysis 1

The following section outlines the results of the analysis following encoding approach 1, that is thresholding CP, STs, and HD, while transforming ECGr, CA, and Thal into ordered factors.

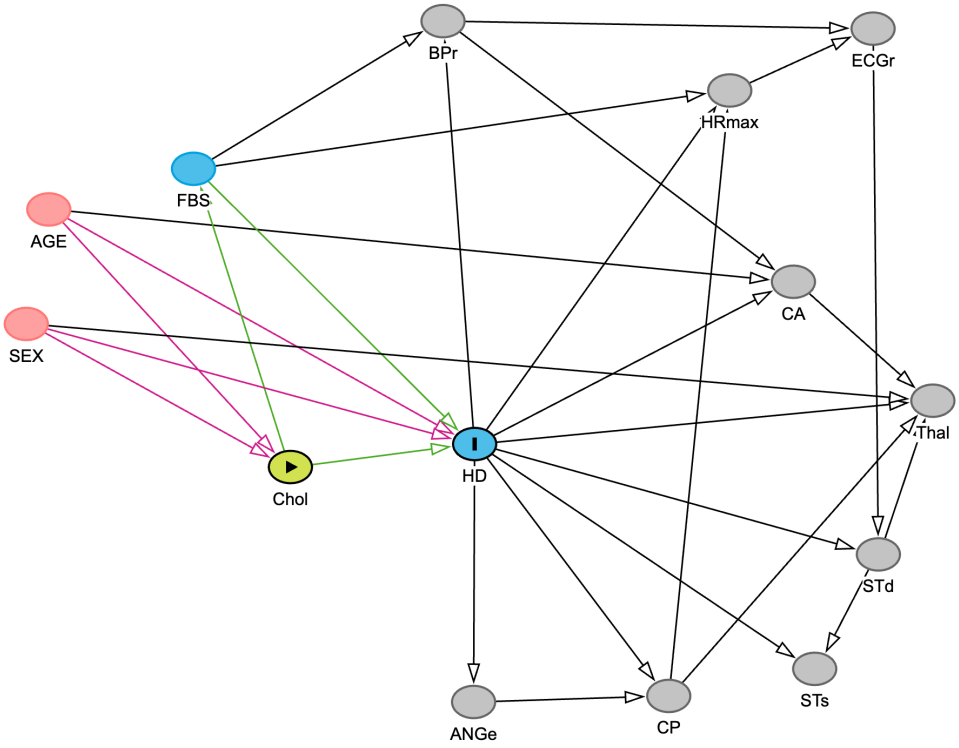


Figure 4: Final DAG from Analysis 1

3.2.1 Model Testing

In Table 4, we show the most egregious of the conditional independence violations. As can be seen, most violations involved either a conditional independence assumption between SEX and Thal, or AGE and CA. Therefore, we dealt with these two issues. We first added an edge between SEX and Thal, confirmed that the violations involving AGE and CA still held, then added an age between AGE and CA. The direction of these edges were from AGE/SEX, and to CA/Thal. The model pruning showed only one edge relationship with a coefficient of absolute value smaller than 0.01. This was the edge from CA to ANGe, with a coefficient of 0.0049. This edge was removed. Our final network for analysis 1 can be seen in Figure 4.

	estimate	p-value	2.5 %	97.5 %
AGE ⫫ CA | FBS, HD	0.4009385	3.916471e-13	0.3006682	0.4947189
SEX ⫫ Thal | CA, CP, ECGr, HD	0.4193977	2.709924e-14	0.3204311	0.5120506
SEX ⫫ Thal | FBS, HD	0.4217148	1.520743e-14	0.3233197	0.5138878
AGE ⫫ CA | BPr, HD	0.4237024	1.099998e-14	0.3254938	0.5157212
SEX ⫫ Thal | CA, CP, HD, STd	0.4277333	7.001323e-15	0.3295535	0.5197366
SEX ⫫ Thal | BPr, CP, ECGr, HD	0.4372212	1.425789e-15	0.3399646	0.5284846
SEX ⫫ Thal | BPr, CP, HD, STd	0.4416928	6.606333e-16	0.3448817	0.5326075
SEX ⫫ Thal | BPr, CP, HD, HRmax	0.4497803	1.591165e-16	0.3537920	0.5400649

Table 4: The worst offenders violating the conditional independence assumptions

3.3 Analysis 2

The following section outlines the results of the analysis following encoding approach 1, that is thresholding CP, STs, and HD, while transforming ECGr, CA, and Thal into ordered factors.

3.3.1 Model Testing

Following the approach outlined in subsection 2.3, the assumptions about conditional independencies made by the DAG structure were tested. This revealed a number of significant violations, the most severe of which related to the relationship

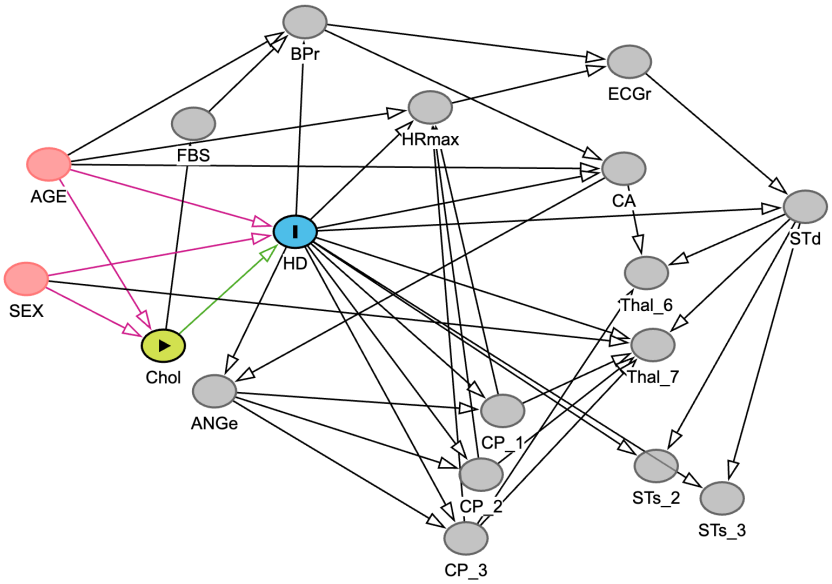


Figure 5: Final DAG from Analysis 2

between AGE and BPr, HRmax, and CA. To adjust for this we added edges between AGE and these variables, and retested the DAG. This showed, that apart from the implied independence of SEX and Thal_7, all violations were found for independencies between different levels of the dummy coded variables (comp. Table 5).

Since adding edges between these levels doesn't add meaningful relationships to the model, we only added an edge from SEX to Thal_7. Lastly, we checked for coefficients within the $[-0.1; 0.1]$ range and found four: CP_1 predicting Thal_6, FBS predicting HD and HRmax, and HD predicting Thal_6. Consequently, we pruned the corresponding edges and finalised the model. The resulting DAG is presented in Figure 5.

Condition	**Estimate**	**p-value**	**2.5%**	**97.5%**
CP_1 ⊥ CP_3 | ANGe, HD	-0.2340368	4.61×10^{-5}	-0.3394820	-0.12313168
CP_2 ⊥ CP_3 | ANGe, HD	-0.4042661	2.37×10^{-13}	-0.4977898	-0.30428700
CP_2 ⊥ STd | ECGr, HD	-0.1936207	8.06×10^{-4}	-0.3013302	-0.08121887
HRmx ⊥ ST_2 | HD, STd	-0.3074888	5.63×10^{-8}	-0.4081789	-0.20034708
HRmx ⊥ ST_2 | ECGr, HD	-0.3292150	5.12×10^{-9}	-0.4283668	-0.22345436
ST_2 ⊥ ST_3 | HD, STd	-0.4236057	1.12×10^{-14}	-0.5156320	-0.32538796
Th_6 ⊥ Th_7 | CA, CP_2, CP_3, STd	-0.2789433	1.06×10^{-6}	-0.3819098	-0.16979208

Table 5: Significant violations of conditional independence assumed by the DAG

3.4 Effect Estimation

The full results for the regression with HD encoded as a binary variable are shown in Table 6. As can be seen, cholesterol was actually the ancestor with the lowest impact, though its effect was still considerable. The coefficient 0.200 can be interpreted as follows. An increase of one standard deviation in the serum cholesterol level causes an increase of odds of having heart disease of factor $e^{0.200} \approx 1.221$.

The impact of fasting blood sugar and sex - both binary variables - were also of a relatively large nature. In our analysis, 0 was coded as female, and 1 as male. These results therefore indicate that for males the odds of having heart disease increase with a factor of $e^{-0.916} \approx 2.5$ compared to otherwise similar females. Moreover, heaving a fasting blood sugar level over 120 mg/dl increased the odds of having heart disease with a factor of approximately 1.859.

With a coefficient of 0.486, however, the numeric Age is the largest contributor to heart disease. Again, this can be

Edge	Adjustment Set	Intercept	Coefficient	Variable
Chol → HD	{AGE, SEX}	-1.876	0.200	Chol
			0.536	AGE
			1.192	SEX
FBS → HD	{Chol}	-1.052	0.620	FBS
			0.162	Chol
AGE → HD	{}	-0.998	0.486	AGE
SEX → HD	{}	-1.609	0.916	SEX

Table 6: Logistic regression coefficients for key edges in the graph, with HD encoded as a binary variable

interpreted as follows: as your age increases by one standard deviation, your log odds of heart disease increases by a factor of 0.486. The odds ratio increases therefore by $e^{0.486} \approx 1.625$. In the original data, the mean age was 54 years, and the standard deviation was 9 years. That is, for someone at 54 years of age, growing older by 9 years increases their odds of being diagnosed with heart disease by 62.5%.

4 Discussion

In this report, we constructed DAGs and tested them based on the Heart Disease dataset, and estimated the effect of the predictors in the DAG on the outcome variable heart disease (HD). Within this analysis, we compared two approaches for encoding some of the categorical variables in the dataset, by thresholding them into binary values versus dummy coding them. This resulted in two finalised DAGs that estimate the causal structure within the dataset, that can be used to interpret the effect of different predictors on heart disease, as well as the extent to which heart disease is reflected within the diagnostic measurements of the dataset. Interestingly, the two approaches resulted in differences between the DAG structures, namely with some of the edges being removed between some but not all of the dummy levels of the dummy coded variables compared to the binary-coded versions of the same variables. This difference supports the use of dummy coding for the categorical variables that can be assumed to be unordered, that is chest pain type (CP), slope of the ST segment (STs), and outcome of the thalach test (Thal). This is similarly exemplified by the assumption violations of independence of CP types conditioned on HD and ANGe - since the difference in types is anginal vs non-anginal, it makes sense, that the information provided by HD and ANGe informs the chest pain type.

Finally, we performed regression analysis of the predictor variables age, sex, serum cholesterol (Chol), and fasting blood sugar (FBS). The findings from this, as summarized in Table 6, highlight the relative contributions of several predictors to the odds of having heart disease when HD is encoded as a binary variable. Cholesterol, though significant, exhibited the lowest impact among the predictors, with a coefficient of 0.2. Conversely, fasting blood sugar and sex, both binary variables, showed relatively large effects. Specifically, based on the regression, both having high fasting blood sugar and being male significantly increases the chances of having heart disease. Finally, the analysis revealed that age is the strongest predictor for heart disease amongst all the predictors. Given that heart disease is increasingly becoming an issue in simultaneously ageing societies, this seems very intuitive. While these results do not support the hypothesis, that serum cholesterol is a strong cause of heart disease, they show that specific risk groups can be easily identified, and suggest that targeted preventions and interventions might be the most important takeaway for clinical approaches.

From a more abstract perspective, the project has illustrated the benefits and pitfalls of different ways of processing mixed data in bayesian network analysis. Both the pre-processing of the data and the construction of our DAG shows that there is no modelling without assumptions. Specifically, our DAG was constructed in the condition - diagnosis - symptom schema. This meant that, due to our construction, the variable of interest - heart diagnosis - was placed at the ”heart” of the graph, and not towards its causal ”tail-end”. Because of this, there were only so many causal relationships of interest to test. In an inference setting, this schema might be more appropriate, but for estimating causal effects, it might have caused us to become tunnel-visioned on the initial set-up of the DAG, and what we had initially categorised as ”symptoms”. It might be the case that some of these symptoms are indeed causally ”before” the heart diagnosis itself, even if they would only be ”measured” or discovered post-diagnosis. Given more time, and arriving upon this consideration earlier, we might have constructed a different DAG.

5 Conclusion

To summarize, within the project we succesfully constructed and adjusted two alternative DAGs to explain the data from the Heart Disease dataset, and estimated the causal effect of the predictors age, sex, serum cholesterol, and fasting blood sugar, on heart disease. The model construction revealed that age has a significant, irreducible effect on a number of diagnostic outcomes, and that these were more nuanced for the unordered categorical outcomes using dummy coding.

The effect estimation revealed that age is the strongest predictor of heart disease, followed by sex and fasting blood sugar, with serum cholesterol having the smallest coefficient.

These results indicate that - holding the magnitude of successful intervention constant - intervening on fasting blood sugar might actually be more effective than intervening on cholesterol. Additionally, interventions should be prioritized - unless there is a non-linear action between age and intervention efficacy - for those of older age.

In the end, we hope that the results created here can be used for a better foundation for modelling and decision-making for heart disease risk, and that these estimates of effect can be used in a pipeline for deciding which interventions to carry out, scale, or prioritize.

References

[1] C. for Disease Control and Prevention, *Heart disease facts*, 2024. [Online]. Available: https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html.

[2] Eurostat, *Cardiovascular diseases statistics - statistics explained*, Jul. 2024. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php/Cardiovascular_diseases_statistics.

[3] S. Berger, G. Raman, R. Vishwanathan, P. F. Jacques, and E. J. Johnson, "Dietary cholesterol and cardiovascular disease: A systematic review and meta-analysis," *The American journal of clinical nutrition*, vol. 102, no. 2, pp. 276–294, 2015.

[4] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, *Heart Disease*, DOI: https://doi.org/10.24432/C52P4X, 1989.

[5] Y. Rosseel, "lavaan: An R package for structural equation modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, 2012. DOI: 10.18637/jss.v048.i02.

[6] J. Textor, B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison, "Robust causal inference using directed acyclic graphs: The r package 'dagitty'," *International Journal of Epidemiology*, vol. 45, no. 6, pp. 1887–1894, 2016. DOI: 10.1093/ije/dyw341.

[7] M. Barrett, *Ggdag: Analyze and create elegant directed acyclic graphs*, R package version 0.2.13.9000, https://r-causal.github.io/ggdag/, 2024. [Online]. Available: https://github.com/r-causal/ggdag.

[8] A. H. Kashou, *St segment*, Aug. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/sites/books/NBK459364/.