Bayesian Networks and Causal Inference Assignment 1

Evander van Wolfswinkel
s1057895

Janneke Verbeek
s1011065

Niek Derksen
s4363779

Introduction

Heart failures, strokes and other types of pathologies belonging to cardiovascular diseases cause the death of approximately 17 million people worldwide annually [13]. Heart failure in particular knows multiple factors that affect the likelihood of occurring. Information extracted from electronic health records can be used to model the relationship between heart failure and risk factors. Due to the broad set of predictors and indicators, modeling heart failure survival is still a hard problem, with limited interpretability of the prediction variables. Most worth noting is the fact that Congestive Heart Failure (CHF) is a chronic and progressive condition that affects the pumping power of the heart muscle, and not a single event. This is why considering causal influence of multiple factors might present a more cohesive insight into an approach to tackling this progressive disease.

Several studies have been conducted on debunking the myths on risk factors. In this project the driving variables of heart failure risks are investigated by employing a Bayesian approach which will serve to investigate causal structure within the data variable space. This will be done by means of fitting a Structural Equation Model (SEM) to data which resulted from a hospital study concerning CHF. This model will be tested and evaluated using widely used discrepancy functions [17]. The evaluation results will be iteratively used to enhance this projects understanding of causal pathways within context of this CHF study. The gained insights will allow for model readjustment, a consequent better model fit and further improving overall understanding of causal pathways within the given context. Lastly, this project will discuss the choices made, overall results and discuss future improvements.

DATA

The data used in this project was collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) and contains medical records of 299 heart failure patients [1]. The data consists of 14 variables, where the variable of interest is $DEATH\_EVENT$, indicating whether a subject has died from CHF. An overview of these variables, their types and their ranges can be found in Table 1.

The data set is distributed as follows: As shown in Figure 1 some variables show a skewed distribution. After closer analysis, some outliers where detected:

- The creatine phosphokinase (CPK) variable has a normal (healthy) value range between 0 and 120 micro grams per liter [11]. Yet, there are cases in which these CPK values reach 7861 mcg/L, which could indicate a measurement of a patient in acute cardiac arrest, but more likely an error. This outlier could cause variance growth issues in the SEM fitting and its removal could be argued. Yet definitive evidence that this measurement did indeed involve an error could not be found, so it was kept within the data set.
- The platelets variable has a normal (healthy) value range between 150.000-450.000 platelets per microliter [12]. Some values ranged upwards of 850.000 platelets per microliter. This could indicate an error, but it is also likely the patient in question suffered from thrombocytosis or a massive inflammatory response within the body.

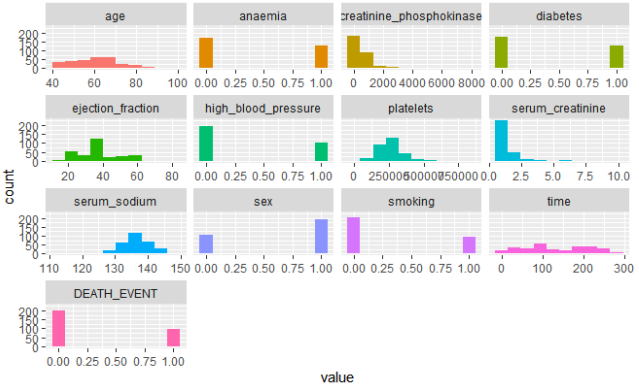

Figure 1: Distribution of variables within the CHF data set.

The variables where subsequently binned, as the outliers causing skewed data distribution for some variables could cause variance explosion within SEM model fitting process. The number of levels each variable has after binning is included after the slash in Table 1. Binning was done by examining the histograms of the data and dividing the data into regularly spaced bins. The above mentioned outliers where grouped in the last bin for each corresponding variable.

STRUCTURAL EQUATION MODELLING

Approach

Before starting on the design of a causal structure, the domain was taken into the equation. Many clinical studies have been conducted within the field of heart disease, and where therefor consulted. This gave a solid base to derive a first basic network from, which is displayed in figure 2.

The network was built using mostly theoretical knowledge,

Table 1: Table of variables in the CHF dataset

Name	Type	Range/levels
age	numeric	40-95 / 6
anaemia	categorical binary	2
creatine phosphokinase	continuous	23-7861 / 11
DEATH_EVENT	categorical binary	2
diabetes	categorical binary	2
ejection fraction	continuous	14-80 / 6
high blood pressure	categorical binary	2
kidney failure	categorical binary	2
platelets	continuous	25100-850000 / 11
serum creatinine	continuous	0.5-9.4 / 8
serum sodium	continuous	113-148 / 6
sex	categorical binary	2
smoking	categorical binary	2
time	numeric	4-285 / 10

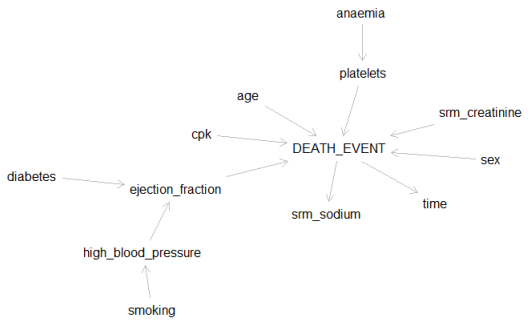

Figure 2: Graph G1; basic causal structure derived from domain knowledge.

in addition to some suspicions of which variables may affect each other. The theoretical knowledge was mostly garnered from literature, as the team does not have a lot of pre-existing domain knowledge. For instance, smoking has been shown to affect blood pressure [6], therefore, an edge between smoking and high blood pressure was drawn. Some relation between the platelet count and anaemia exists [7], so there is an edge drawn there as well. In patients with high blood pressure and reduced ejection fraction, there appears to be a higher heart failure survival rate.

There is some literature suggesting that there may be latent variables present in the network. For instance, diabetes and high blood pressure could be related through a hidden obesity variable. Furthermore, serum creatinine is associated with kidney failure, which in turn can also be associated with anaemia and congestive heart failure, which could be a latent variable. However, we omit these latent associations from our first model, as we needed this first approach to serve as a benchmark. Any future model based on and adapted from this benchmark, should thus reliably inform us on the fit improvement of any made changes based on iterative testing described later in this project.

Implementation

The network was implemented in R, mainly using the `lavaan` package in order to fit SEM models, study the results and plot their corresponding coefficients. For graphical representation of models this project used `dagitty` [16], specifically the web-applet. Some convenient functions from the `bnlearn` package have also been used, such as `localTests` and `toString`.

TESTS

By performing independence tests on the entire structure, we can examine whether our network structure has been specified in a sensible way. For this purpose we used a combination of chi-squared tests (as all continuous variables have been binned) and polychoric correlation, which works with both categorical and continuous variables by assuming all binned variables are derived from latent normally distributed continuous data. Performing these test iteratively while challenging the results against the known scientific theories will help derive a quasi-final model.

First iteration testing

Results. Chi-squared and polychoric correlation test statistics were computed for implied independent relationships of the model. The 12 results with the lowest p-values for the chi-squared test can be seen in Table 2.

Table 2: Table of top chi-squared implied dependent relations ordered by p-value, test 1

	rmsea	x2	df	p.value
sex ⫫ smkn	0.44	59.45	1.00	0.00
cpk ⫫ srm_sod | DEATH	0.05	192.72	90.00	0.00
cpk ⫫ time | DEATH	0.07	248.38	154.00	0.00
srm_c ⫫ srm_sod | DEATH	0.07	126.43	70.00	0.00
age ⫫ srm_c	0.05	65.41	35.00	0.00
anam ⫫ cpk	0.08	27.19	10.00	0.00
dbts ⫫ sex	0.15	7.44	1.00	0.01
dbts ⫫ smkn	0.14	6.48	1.00	0.01
age ⫫ hbp	0.08	13.60	5.00	0.02
hbp ⫫ time | ejc_frac	0.13	66.95	45.00	0.02
age ⫫ anam	0.06	10.80	5.00	0.06
DEATH ⫫ hbp | ejc_frac	0.17	11.94	6.00	0.06

cpk → creatine phosphokinase
hbp → high blood pressure
srm_c and srm_sod → serum creatinine, serum sodium
ejc_frac → ejection fraction
anam, smkn, dbts → anaemia, smoking, diabetes

Misspecifications. From the chi-squared test on this network specification, it seems we have neglected to model relationships between sex and a number of variables. The independence test for sex ⫫ smoking failed ($p < .05$, $RMSEA = 0.44$). Other implied significant dependencies from the chi-squared test include diabetes dependent on both sex and smoking ($p < .05$ and $p < .05$ respectively, $RMSEA$ 0.14). When looking

at the polychoric correlation tests, estimation coefficients suggests the same dependent relations as for the chi-squared test, such as the relationship between smoking, sex and diabetes.

Literature investigation. When investigating these implied dependencies, we challenged them against known scientific relationships found in literature. The connection for smoking and sex is well-documented in literature. According to a 2014 survey Pakistani men smoke significantly more than women (31.8% of males and 5.8% of females reported current smoking of tobacco) [14]. The relationship implied for diabetes and sex was not found in a 2017 survey of 10800 Pakistani, but concluded a large difference of obesity between men and women living in rural or urban locations, suggesting a possible latent variable [2].

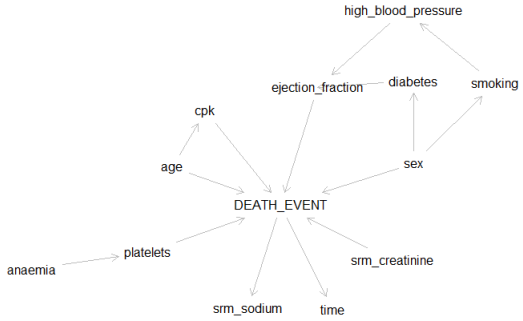

Figure 3: Intermediate model changes based on implied dependencies involving the sex variable

Model changes. To resolve the misspecifications involving sex in one go, we first draw a new edge from sex to smoking as suggested by the literature investigation and the high RMSEA. This introduces some new implied independencies in the network, in particular ejection fraction $\perp\!\!\!\perp$ sex | smoking and both diabetes $\perp\!\!\!\perp$ sex and diabetes $\perp\!\!\!\perp$ smoking. Since the literature investigation suggests that sex and diabetes are in fact dependent, we also draw an edge from sex to diabetes, such that diabetes $\not\perp\!\!\!\perp$ sex, and diabetes $\perp\!\!\!\perp$ smoking | sex. These intermediate model changes can be seen in Figure 3.

Second iteration testing

Results. Chi-squared and polychoric correlation test statistics were computed for implied independent relationships of the model. The 12 results with the lowest p-values for the chi-squared test can be seen in Table 3

From the second test, we can observe that the implied independence diabetes $\perp\!\!\!\perp$ smoking | sex does not have a significant p-value, suggesting that the change made to the model after the previous test was somewhat effective.

Now, the implied dependencies exhibiting the highest RMSEAs are the following. The test for $DEATH\_EVENT \perp\!\!\!\perp$ high blood pressure | ejection fraction, smoking ($p < .05$, *RMSEA* 0.26) failed, as well as high blood pressure $\perp\!\!\!\perp$ time | ejection fraction, diabetes, smoking ($p < .05$, *RMSEA* 0.24).

Table 3: Table of top chi-squared implied dependent relations ordered by p-value, test 2

	rmsea	df	p.value
cpk $\perp\!\!\!\perp$ srm_sod | DEAT	0.05	90.00	0.00
cpk $\perp\!\!\!\perp$ time | DEAT	0.07	154.00	0.00
srm_c $\perp\!\!\!\perp$ srm_sod | DEAT	0.07	70.00	0.00
age $\perp\!\!\!\perp$ srm_sod	0.05	35.00	0.00
DEAT $\perp\!\!\!\perp$ hbp | dbts, ejc_frac, smkn	0.26	15.00	0.00
anam $\perp\!\!\!\perp$ cpk	0.08	10.00	0.00
DEAT $\perp\!\!\!\perp$ smkn | ejc_frac, sex	0.29	8.00	0.01
age $\perp\!\!\!\perp$ hbp	0.08	5.00	0.02
hbp $\perp\!\!\!\perp$ time | dbts, ejc_frac, smkn	0.24	99.00	0.03
ejc $\perp\!\!\!\perp$ sex | dbts, hbp	0.07	19.00	0.05
age $\perp\!\!\!\perp$ anam	0.06	5.00	0.06
anam $\perp\!\!\!\perp$ smkn	0.09	1.00	0.06

cpk → creatine phosphokinase
hbp → high blood pressure
srm_c and srm_sod → serum creatinine, serum sodium
ejc_frac → ejection fraction
anam, smkn, dbts → anaemia, smoking, diabetes

Misspecifications. These results suggest that there are missing dependencies in our model, meaning that there should be no independence between high blood pressure and death as well as high blood pressure and time when conditioning on ejection fraction, smoking, and in case of time, diabetes.

Literature investigation. High blood pressure has been shown to increase the risk of death from congestive heart failure. Chael et al. found that in the elderly, with every 10-mm Hg rise in systolic blood pressure, there was a 12% increase in risk of CHF [4]. Haider et al. found that increases in systolic pressure and pulse pressure are especially associated with an increased risk of CHF, even when controlling for age [8].

Model changes. The simplest way to resolve the misspecification is by drawing an edge between high blood pressure and death, as this would make both time and death dependent on high blood pressure given ejection fraction, smoking and diabetes. The final model can be seen in Figure 4.

APPLICATION

Method. In order to examine the causal effects of different variables in our model on each other, we fitted our tested model using `lavaan` function `sem` and examined the path coefficients (as association measures derived from polychoric correlation) between different variables. The resulting largest coefficients are plotted in Figure 5.

Results. From the graph in Figure 5 it can be seen that the largest effect is that of $DEATH\_EVENT$ and time. $DEATH\_EVENT$ is positive when a patient died during the period defined in time (follow-up). This suggest patients are more likely to decease when the follow up period is short.

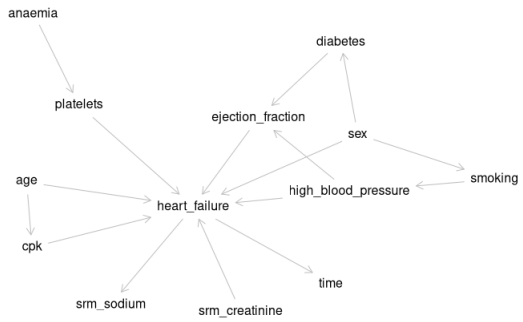

Figure 4: Final model

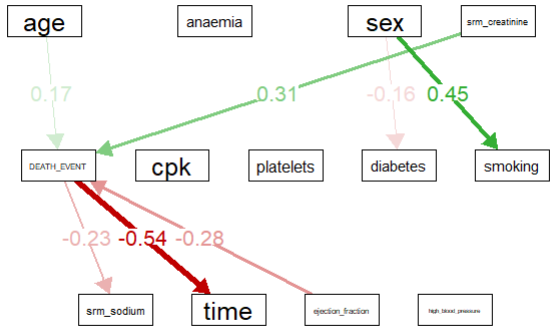

Figure 5: Graph of the edge coefficients computed using polychoric correlation by `lavaan`

This shorter follow up period could signify that patients which are in a critical condition require more medical attention, resulting in a larger portion of fatalities. Another large positive association is observed for serum creatinine levels and $DEATH\_EVENT$, the increase in serum creatinine associated with mortality is well documented in literature [3], but is also known to be dependent on other variables such as age [15] and gender [10]. Other interesting path associations include the decrease in serum sodium with $DEATH\_EVENT$, a association also found in retrospective analysis studies [9], and the association for ejection fraction and $DEATH\_EVENT$. Research suggest that reduced ejection fraction increases the chance of mortality while preserved ejection fraction improves chances [5].

DISCUSSION

Since a model that captures all dependencies in the data perfectly may overfit, the model was not changed further after the largest issues with it were resolved, as described above. Thus, in the final model, there are still some implied independencies that fail a chi-squared test, but most of these independencies have a RMSEA of 0.05-0.07, such as CPK $\perp\!\!\!\perp$ serum sodium $\mid DEATH\_EVENT$ ($p < .001$, $RMSEA$ 0.05). Since the continuous variables were binned, some information was lost there.

Though initially the suspicion was that there could be latent variables in the model, the largest issues in the model did not require us to introduce them. Since we did not attempt to resolve all missing dependencies in our model, it still could be the case that further model changes would require us to introduce a latent variable. Further tests would be needed to determine whether this is indeed the case.

Determining the path coefficients was a fairly smooth process, as the polychoric correlation method used is quite practical when continuous variables are binned, as values can then be conveniently assumed to be rested on an underlying joint continuous normal distribution. The assumption that this underlying distribution is normal does however make this method less statistically robust.

REFERENCES

[1] Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza. 2017. Survival analysis of heart failure patients: A case study. *PloS one* 12, 7 (2017), e0181001.

[2] Abdul Basit, Asher Fawwad, Huma Qureshi, and AS Shera. 2018. Prevalence of diabetes, pre-diabetes and associated risk factors: second National Diabetes Survey of Pakistan (NDSP), 2016–2017. *BMJ open* 8, 8 (2018), e020961.

[3] Javed Butler, Diana Chirovsky, Hemant Phatak, Anne McNeill, and Robert Cody. 2010. Renal function, health outcomes, and resource utilization in acute heart failure: a systematic review. *Circulation: Heart Failure* 3, 6 (2010), 726–745.

[4] Claudia U Chae, Marc A Pfeffer, Robert J Glynn, Gary F Mitchell, James O Taylor, and Charles H Hennekens. 1999. Increased pulse pressure and risk of heart failure in the elderly. *Jama* 281, 7 (1999), 634–643.

[5] Shannon M Dunlay, Véronique L Roger, Susan A Weston, Ruoxiang Jiang, and Margaret M Redfield. 2012. Longitudinal changes in ejection fraction in heart failure patients with preserved and reduced ejection fraction. *Circulation: Heart Failure* 5, 6 (2012), 720–726.

[6] Antonella Groppelli, DM Giorgi, Stefano Omboni, Gianfranco Parati, and Giuseppe Mancia. 1992. Persistent blood pressure increase induced by heavy smoking. *J hypertens* 10, 5 (1992), 495–499.

[7] Samuel Gross, Vicki Keefer, and Arthur J Newman. 1964. The platelets in iron-deficiency anemia. I. The response to oral and parenteral iron. *Pediatrics* 34, 3 (1964), 315–323.

[8] Agha W Haider, Martin G Larson, Stanley S Franklin, and Daniel Levy. 2003. Systolic blood pressure, diastolic blood pressure, and pulse pressure as predictors of risk for congestive heart failure in the Framingham Heart Study. *Annals of internal medicine* 138, 1 (2003), 10–16.

[9] Liviu Klein, Christopher M O'Connor, Jeffrey D Leimberger, Wendy Gattis-Stough, Ileana L Piña, G Michael Felker, Kirkwood F Adams Jr, Robert M Califf, and Mihai Gheorghiade. 2005. Lower serum sodium is associated with increased short-term mortality in hospitalized patients with worsening heart failure: results from the Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure (OPTIME-CHF) study. *Circulation* 111, 19 (2005), 2454–2460.

[10] Harlan M Krumholz, Ya-Ting Chen, Viola Vaccarino, Yun Wang, Martha J Radford, W David Bradford, and Ralph I Horwitz. 2000. Correlates and impact on outcomes of worsening renal function in patients $\geq$ 65 years of age with heart failure. *The American journal of cardiology* 85, 9 (2000), 1110–1113.

[11] Richard A McPherson, MD Msc, and Matthew R Pincus. 2021. *Henry's clinical diagnosis and management by laboratory methods E-book*. Elsevier Health Sciences.

[12] Dennis W Ross, Lanier H Ayscue, Judith Watson, and Stuart A Bentley. 1988. Stability of hematologic parameters in healthy subjects: Intraindividual versus interindividual variation. *American journal of clinical pathology* 90, 3 (1988), 262–267.

[13] Gregory A Roth, George A Mensah, Catherine O Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M Baddour, Noël C Barengo, Andrea Z Beaton, Emelia J Benjamin, Catherine P Benziger, et al. 2020. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *Journal of the American College of Cardiology* 76, 25 (2020), 2982–3021.

[14] Muhammad Arif Nadeem Saqib, Ibrar Rafique, Huma Qureshi, Muhammad Arif Munir, Rizwan Bashir, Babur Wasim Arif, Khalid Bhatti, Shahzad Alam Khan Ahmed, and Lubna Bhatti. 2018. Burden of tobacco in Pakistan: findings from global adult tobacco survey 2014. *Nicotine and Tobacco Research* 20, 9 (2018), 1138–1143.

[15] Jeffrey M Testani, Steven G Coca, Brian D McCauley, Richard P Shannon, and Stephen E Kimmel. 2011. Impact of changes in blood pressure during the treatment of acute decompensated heart failure on renal and clinical outcomes. *European journal of heart failure* 13, 8 (2011), 877–884.

[16] Johannes Textor, Benito van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. 2016. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International journal of epidemiology* 45, 6 (2016), 1887–1894.

[17] Yan Xia and Yanyun Yang. 2019. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior research methods* 51, 1 (2019), 409–428.

APPENDIX

Import packages

```
library(semPlot)
library(bayesianNetworks)
library(funModeling)
library(tidyverse)
library(Hmisc)
library(bnlearn)
library(naivebayes)
library(dagitty)
library(dataPreparation)
library(lavaan)
library(summarytools)
library(OneR)
library(corrplot)
library(knitr)
library(readxl)
library(dplyr)
library(kableExtra)
```

```
d1=read.table("./heart_failure_clinical_records_dataset.csv", sep=',', header=TRUE)
```

```
dfSummary(d1, plain.ascii  = FALSE,
          style        = 'grid',
          graph.magnif = 0.85,
          varnumbers = FALSE,
          valid.col    = FALSE, tmp.img.dir = "/tmp")
```

```
describe(d1)  %>% html()
```

```
plot_num(d1)
```

Preprocessing

Shorten some long variable names.

```
d1 <- rename(d1, cpk = creatinine_phosphokinase)
d1 <- rename(d1, srm_creatinine = serum_creatinine)
d1 <- rename(d1, srm_sodium = serum_sodium)
```

Create extra dataset to perform processing on.

```
d1_proc = data.frame(d1)
```

Continuous and binary data. Order binary variables

```
d1_proc$anaemia <- as.numeric(ordered(d1$anaemia))
d1_proc$sex <- as.numeric(ordered(d1$sex))
d1_proc$high_blood_pressure <- as.numeric(ordered(d1$high_blood_pressure))
d1_proc$diabetes <- as.numeric(ordered(d1$diabetes))
d1_proc$smoking <- as.numeric(ordered(d1$smoking))
d1_proc$DEATH_EVENT <- as.numeric(ordered(d1$DEATH_EVENT))
```

Bin all continuous. Bin continuous variables using the histograms of the data to create regularly spaces bins

```
temp_age <- rep("<97", nrow(d1_proc)) #96 is max age
temp_age[d1_proc$age >=0 & d1_proc$age <50] <- "<50"
temp_age[d1_proc$age >=50 & d1_proc$age <60] <- "50-60"
temp_age[d1_proc$age >=60 & d1_proc$age <70] <- "60-70"
temp_age[d1_proc$age >=70 & d1_proc$age <80] <- "70-80"
temp_age[d1_proc$age >=80 & d1_proc$age<90] <- "80-90"
```

```
temp_age[d1_proc$age >=90] <- "90>"
#turn binned data into factor:
d1_proc$age <- ordered(temp_age, levels=c("<50","50-60","60-70","70-80","80-90",
                                          "90>"))

temp_cpk <- rep("<7862", nrow(d1_proc)) #96 is max cpk
temp_cpk[d1_proc$cpk >=0 & d1_proc$cpk <200] <- "0-200"
temp_cpk[d1_proc$cpk >=200 & d1_proc$cpk <400] <- "200-400"
temp_cpk[d1_proc$cpk >=400 & d1_proc$cpk <600] <- "400-600"
temp_cpk[d1_proc$cpk >=600 & d1_proc$cpk <800] <- "600-800"
temp_cpk[d1_proc$cpk >=800 & d1_proc$cpk <1000] <- "800-1000"
temp_cpk[d1_proc$cpk >=1000 & d1_proc$cpk <1200] <- "1000-1200"
temp_cpk[d1_proc$cpk >=1200 & d1_proc$cpk <1400] <- "1200-1400"
temp_cpk[d1_proc$cpk >=1400 & d1_proc$cpk <1600] <- "1400-1600"
temp_cpk[d1_proc$cpk >=1600 & d1_proc$cpk <1800] <- "1600-1800"
temp_cpk[d1_proc$cpk >=1800 & d1_proc$cpk <2000] <- "1800-2000"
temp_cpk[d1_proc$cpk >=2000] <- "2000>"
#turn binned data into factor:
d1_proc$cpk <- ordered(temp_cpk, levels=c("0-200", "200-400", "400-600",
                                          "600-800","800-1000","1000-1200",
                                          "1200-1400","1400-1600",
                                          "1600-1800","1800-2000",
                                          "2000>"))

temp_ef <- rep("<81", nrow(d1_proc)) #80 is max ef
temp_ef[d1_proc$ejection_fraction >=0  & d1_proc$ejection_fraction <20] <- "<20"
temp_ef[d1_proc$ejection_fraction >=20 & d1_proc$ejection_fraction <30] <- "20-30"
temp_ef[d1_proc$ejection_fraction >=30 & d1_proc$ejection_fraction <40] <- "30-40"
temp_ef[d1_proc$ejection_fraction >=40 & d1_proc$ejection_fraction <50] <- "40-50"
temp_ef[d1_proc$ejection_fraction >=50 & d1_proc$ejection_fraction <60] <- "50-60"
temp_ef[d1_proc$ejection_fraction >=60] <- "60>"
#turn binned data into factor:
d1_proc$ejection_fraction <- ordered(temp_ef, levels=c("<20","20-30","30-40",
                                                       "40-50","50-60","60>"))

temp_plt <- rep("<850k", nrow(d1_proc)) #850k is max platelets
temp_plt[d1_proc$platelets >=0  & d1_proc$platelets <50000] <- "<50k"
temp_plt[d1_proc$platelets >=50000  & d1_proc$platelets <100000] <- "50k-100k"
temp_plt[d1_proc$platelets >=100000  & d1_proc$platelets <150000] <- "100k-150k"
temp_plt[d1_proc$platelets >=150000  & d1_proc$platelets <200000] <- "150k-200k"
temp_plt[d1_proc$platelets >=200000  & d1_proc$platelets <250000] <- "200k-250k"
temp_plt[d1_proc$platelets >=250000  & d1_proc$platelets <300000] <- "250k-300k"
temp_plt[d1_proc$platelets >=300000  & d1_proc$platelets <350000] <- "300k-350k"
temp_plt[d1_proc$platelets >=350000  & d1_proc$platelets <400000] <- "350k-400k"
temp_plt[d1_proc$platelets >=400000  & d1_proc$platelets <450000] <- "400k-450k"
temp_plt[d1_proc$platelets >=450000  & d1_proc$platelets <500000] <- "450k-500k"
temp_plt[d1_proc$platelets >=500000] <- "500k>"
#turn binned data into factor
d1_proc$platelets <- ordered(temp_plt, levels=c("<50k","50k-100k","100k-150k",
                                                "150k-200k","200k-250k","250k-300k",
                                                "300k-350k","350k-400k",
                                                "400k-450k","450k-500k","500k>"))

temp_sc <- rep("<9.5", nrow(d1_proc)) #850k is max platelets
temp_sc[d1_proc$srm_creatinine >=0  & d1_proc$srm_creatinine <1.0] <- "<1.0"
temp_sc[d1_proc$srm_creatinine >=1.0  & d1_proc$srm_creatinine <1.5] <- "1.0-1.5"
temp_sc[d1_proc$srm_creatinine >=1.5  & d1_proc$srm_creatinine <2.0] <- "1.5-2.0"
temp_sc[d1_proc$srm_creatinine >=2.0  & d1_proc$srm_creatinine <2.5] <- "2.0-2.5"
temp_sc[d1_proc$srm_creatinine >=2.5  & d1_proc$srm_creatinine <3.0] <- "2.5-3.0"
temp_sc[d1_proc$srm_creatinine >=3.0  & d1_proc$srm_creatinine <3.5] <- "3.0-3.5"
temp_sc[d1_proc$srm_creatinine >=3.5  & d1_proc$srm_creatinine <4.0] <- "3.5-4.0"
temp_sc[d1_proc$srm_creatinine >=4.0] <- "4.0>"
d1_proc$srm_creatinine <- ordered(temp_sc, levels=c("<1.0","1.0-1.5","1.5-2.0",
                                                    "2.0-2.5","2.5-3.0","3.0-3.5","3.5-4.0","4.0>"))

temp_ss <- rep("<149", nrow(d1_proc)) #148 is max serum_sodium
temp_ss[d1_proc$srm_sodium >=0  & d1_proc$srm_sodium <125] <- "<125"
temp_ss[d1_proc$srm_sodium >=125  & d1_proc$srm_sodium <130] <- "125-130"
temp_ss[d1_proc$srm_sodium >=130  & d1_proc$srm_sodium <135] <- "130-135"
temp_ss[d1_proc$srm_sodium >=135  & d1_proc$srm_sodium <140] <- "135-140"
temp_ss[d1_proc$srm_sodium >=140  & d1_proc$srm_sodium <145] <- "140-145"
temp_ss[d1_proc$srm_sodium >=145] <- "145>"
d1_proc$srm_sodium <- ordered(temp_ss, levels=c("<125","125-130", "130-135",
                                                "135-140","140-145","145>"))

temp_time <- rep("<286", nrow(d1_proc)) #285 is max time
temp_time[d1_proc$time >=0  & d1_proc$time <30] <- "<30"
temp_time[d1_proc$time >=30  & d1_proc$time <60] <- "30-60"
temp_time[d1_proc$time >=60  & d1_proc$time <90] <- "60-90"
temp_time[d1_proc$time >=90  & d1_proc$time <120] <- "90-120"
temp_time[d1_proc$time >=120  & d1_proc$time <150] <- "120-150"
temp_time[d1_proc$time >=150  & d1_proc$time <180] <- "150-180"
temp_time[d1_proc$time >=180  & d1_proc$time <210] <- "180-210"
temp_time[d1_proc$time >=210  & d1_proc$time <240] <- "210-240"
temp_time[d1_proc$time >=240  & d1_proc$time <270] <- "240-270"
temp_time[d1_proc$time >=270] <- "270>"
d1_proc$time <- ordered(temp_time,  levels=c("<30","30-60","60-90","90-120",
```

```
                                    "120-150","150-180","180-210",
                                    "210-240","240-270","270>"))
```

Define first model

```
g1 <- graphLayout(dagitty('dag {
bb="0,0,1,1"
age [pos="0.213,0.767"]
anaemia [pos="0.913,0.191"]
cpk [pos="0.814,0.701"]
diabetes [pos="0.387,0.127"]
ejection_fraction [pos="0.526,0.237"]
DEATH_EVENT [pos="0.504,0.491"]
high_blood_pressure [pos="0.661,0.123"]
platelets [pos="0.761,0.379"]
sex [pos="0.385,0.926"]
smoking [pos="0.523,0.027"]
srm_creatinine [pos="0.677,0.933"]
srm_sodium [pos="0.128,0.495"]
time [pos="0.196,0.265"]
age -> DEATH_EVENT
anaemia -> platelets
cpk -> DEATH_EVENT
diabetes -> ejection_fraction
ejection_fraction -> DEATH_EVENT
DEATH_EVENT -> srm_sodium
DEATH_EVENT -> time
high_blood_pressure -> ejection_fraction
platelets -> DEATH_EVENT
sex -> DEATH_EVENT
smoking -> high_blood_pressure
srm_creatinine -> DEATH_EVENT
}
'))
```

```
chi_square_test <- localTests(g1, d1_proc, type = 'cis.chisq')
top_rmsea <- chi_square_test[order(chi_square_test$p.value, decreasing = FALSE),]
knitr::kable(top_rmsea[1:10,1:4])%>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),latex_options = "HOLD_position")
```

```
# Compute polychoric correlation
d1_proc_corr = lavCor(d1_proc)
```

```
corrtest <- localTests(g1, sample.cov = d1_proc_corr, sample.nobs=nrow(d1_proc))
top_corr <- corrtest[order(corrtest$estimate,decreasing = TRUE),]
knitr::kable(top_corr[1:6,1:4])%>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),latex_options = "HOLD_position")
```

```
down_corr <- corrtest[order(corrtest$estimate,decreasing = FALSE),]
knitr::kable(down_corr[1:6,1:4])%>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),latex_options = "HOLD_position")
```

Define intermediate model with sex based changes

```
g2 <- graphLayout(dagitty('dag {
bb="0,0,1,1"
age [pos="0.213,0.767"]
```

```
anaemia [pos="0.913,0.191"]
cpk [pos="0.814,0.701"]
diabetes [pos="0.387,0.127"]
ejection_fraction [pos="0.526,0.237"]
DEATH_EVENT [pos="0.504,0.491"]
high_blood_pressure [pos="0.661,0.123"]
platelets [pos="0.761,0.379"]
sex [pos="0.385,0.926"]
smoking [pos="0.523,0.027"]
srm_creatinine [pos="0.677,0.933"]
srm_sodium [pos="0.128,0.495"]
time [pos="0.196,0.265"]
age -> DEATH_EVENT
age -> cpk
cpk -> DEATH_EVENT
diabetes -> ejection_fraction
ejection_fraction -> DEATH_EVENT
DEATH_EVENT -> srm_sodium
DEATH_EVENT -> time
high_blood_pressure -> ejection_fraction
anaemia -> platelets
platelets -> DEATH_EVENT
sex -> DEATH_EVENT
smoking -> high_blood_pressure
srm_creatinine -> DEATH_EVENT
sex -> smoking
sex -> diabetes
}
'))
```

Define final model

```
g3 <- graphLayout(dagitty('dag {
bb="0,0,1,1"
age [pos="0.213,0.767"]
anaemia [pos="0.913,0.191"]
cpk [pos="0.814,0.701"]
diabetes [pos="0.387,0.127"]
ejection_fraction [pos="0.526,0.237"]
DEATH_EVENT [pos="0.504,0.491"]
high_blood_pressure [pos="0.661,0.123"]
platelets [pos="0.761,0.379"]
sex [pos="0.385,0.926"]
smoking [pos="0.523,0.027"]
srm_creatinine [pos="0.677,0.933"]
srm_sodium [pos="0.128,0.495"]
time [pos="0.196,0.265"]
age -> DEATH_EVENT
age -> cpk
cpk -> DEATH_EVENT
diabetes -> ejection_fraction
ejection_fraction -> DEATH_EVENT
DEATH_EVENT -> srm_sodium
DEATH_EVENT -> time
high_blood_pressure -> ejection_fraction
anaemia -> platelets
```

```
platelets -> DEATH_EVENT
sex -> DEATH_EVENT
smoking -> high_blood_pressure
srm_creatinine -> DEATH_EVENT
sex -> smoking
sex -> diabetes
high_blood_pressure -> DEATH_EVENT
}
'))
```

Fitting SEM using binned catagorical polychoric correlation matrix

```
# Define SEM model in lavaan syntax
sem_model <- "
            srm_sodium~DEATH_EVENT
            time~DEATH_EVENT
            DEATH_EVENT~age
            cpk~age
            platelets~anaemia
            DEATH_EVENT~cpk
            ejection_fraction~diabetes
            DEATH_EVENT~ejection_fraction
            DEATH_EVENT~high_blood_pressure
            ejection_fraction~high_blood_pressure
            DEATH_EVENT~platelets
            DEATH_EVENT~sex
            diabetes~sex
            smoking~sex
            high_blood_pressure~smoking
            DEATH_EVENT~srm_creatinine
           "

# Fit SEM
fit <- sem(sem_model, sample.cov = d1_proc_corr, sample.nobs = nrow(d1_proc), fixed.x = FALSE)

# Plot SEM network without exogenous covariances, minimum coef values of 0.1 and no residuals.
semPaths(fit, what="est", whatLabels = "par", style = "OpenMx", layout = "tree2",
         residuals = FALSE, nCharNodes=0, edge.label.cex = 1.5, asize = 6,
         sizeMan = 12,sizeMan2 = 5,minimum = 0.1,curvature = 1.5,
         rotation=1,curve=2, exoCov=FALSE)
```