The School of Mathematics

THE UNIVERSITY
of EDINBURGH

Dementia: Investigating Risk Factors by Bayesian Network

by

Haozhe Wang, S2166258

Dissertation Presented for the Degree of
MSc in Statistics with Data Science

July 2022

Supervised by
Dr Sara Wade, Dr Cecilia Balocchi and Steven Soutar

Executive Summary

People with dementia increase with time, so researchers are interested in exploring the risk factor of it to prevent it. Livingston et al. [9] mentioned 12 risk factors related to dementia. However, dementia related to these risk factors may not imply a direct relationship. Therefore, this report will explore the specific relationship of the risk factors with dementia by the Bayesian network, especially sex, age, and country. Nevertheless, the data provided by SHARE does not include all the 12 risk factors. Thus, the direct risk factors learned by the Bayesian network may not be accurate, but it does show the other risk factors are not directly related to dementia. Eventually, the direct risk factors learned from this report are education, depression or activity, and age. Sex and country do not directly relate to dementia.

Acknowledgments

I am grateful to the supervisors of this project, Dr. Sara Wade, Dr. Cecilia Balocchi, and Steven Soutar, for their support and helpful suggestions. I am particularly indebted to Mr. Steven's patience in listening and responding to my questions. By the way, it is grateful for the background of dementia provided by Prof Graciela Muniz-Terrara and Dr. Anja Leist. In the end, I would also like to thank the Survey of Health, Ageing and Retirement in Europe (SHARE) for providing the data and guide for explanation.

Word Counts

4979 words (including the executive summary, main text, references, and appendices). Excluding the acknowledgments, word counts, own work deceleration, contents, list of tables, and list of figures. The screenshot provided in the appendix D.

University of Edinburgh – Own Work Declaration

This sheet must be filled in, signed and dated - your work will not be marked unless this is done.

Name: Haozhe Wang
Matriculation Number: S2166258
Title of work: Dementia: Investigating Risk Factors by Bayesian Network

I confirm that all this work is my own except where indicated, and that I have:

- Clearly referenced/listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Not sought or used the help of any external professional academic agencies for the work
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Complied with any other plagiarism criteria specified in the Course handbook

I understand that any false claim for this work will be penalised in accordance with the University regulations (https://teaching.maths.ed.ac.uk/main/msc-students/msc-programmes/statistics/data-science/assessment/academic-misconduct).

Signature .....

Date .....

Contents

1 Introduction **1**
- 1.1 Background & Motivation . . . 1
- 1.2 Data . . . 1

2 Data pre-processing **1**
- 2.1 Variable selection . . . 1
- 2.2 Data manipulation . . . 1
- 2.3 Multiple Imputation . . . 3

3 Methods for Bayesian Network **5**

4 Bayesian Network workflow **5**
- 4.1 Structural learning part . . . 5
- 4.2 Fit model part . . . 6
- 4.3 Present the estimates . . . 6
- 4.4 Diagnosis . . . 6

5 Results **6**
- 5.1 wave 4 & Estonia . . . 6
- 5.2 multiple waves & multiple countries . . . 11

6 Conclusion **18**

Appendices **22**

A the code **22**

B Appendix for EDA **22**

C Appendix for missing values **23**

D Appendix for word count **23**

List of Tables

1 Variables information . . . 2
2 bic network score for different graphical structure from different algorithm with different data sets (wave 4 Estonia) . . . 7
3 bde network score for different graphical structure from different algorithm with different data sets (wave 4 Estonia) . . . 8
4 log likelihood network score for different graphical structure from different algorithm with different data sets (wave 4 Estonia) . . . 8
5 compare two graph by bn.cv (wave 4 Estonia) . . . 8
6 compare graphs add edges by bn.cv (wave 4 & Estonia) . . . 9
7 bic network score for different graphical structure from different algorithm with different data sets (multiple waves & multiple countries) . . . 12
8 bde network score for different graphical structure from different algorithm with different data sets (multiple waves & multiple countries) . . . 13
9 log likelihood network score for different graphical structure from different algorithm with different data sets (multiple waves & multiple countries) . . . 13
10 compare two graph by bn.cv (multiple waves & multiple countries) . . . 13
11 compare graphs add edges by bn.cv (multiple waves & multiple countries) . . . 14

List of Figures

1 Compare observed values and missing values for the cognitive score. The left one is for wave 4 study. The right one is for multiple wave study. By the way, blue curve is the observed values, red curves are 5 different imputations' plots. . . . 3
2 Compare observed values and missing values for categorical variables. The left one is for all missing category variables in the wave 4 study, the right one is for all missing category variables in the multiple wave study. The variables are NOD, DS, BMI, SMK, DK, VA, HAM, EDUL, and AD (from left to right and top to bottom). By the way, the black plots are the observed plot and the blue plots are 5 different imputations' plots. 4
3 Distribution of age. Two plot on the left are observed distribution of age, the right are missing distribution of age. . . . 4
4 left is dag 18 & right is better dag . . . 9
5 relation of AD . . . 10
6 Markov blanket variables of AD . . . 11
7 diagnosis_1 . . . 11
8 left is dag 14 & right is better dag . . . 15
9 relation of AD . . . 15
10 Markov blanket variables of AD_1. Variables for DS are HAM, NOD and Sex (Top to bottom). Variable for Age is country. . . . 16
11 Markov blanket variables of AD_2. Variables for EDUL are country, Age, and Sex (Top to bottom) . . . 17
12 diagnosis_2 . . . 18
13 this is the plot of variables before EDA. . . . 22
14 this is the plot of variables after EDA. . . . 23
15 missing pattern for wave 1 (left) and 2 (right). Both plots are point that combine_cognitive is complete missing. . . . 24
16 missing pattern for wave 4 (left) and 5 (right). Both plots are point that no variable is complete missing. The wave is a little better than wave 5. . . . 24
17 missing pattern for wave 6 (left), 7 (right) and 8 (bottom). All plots are point that drinking_behavior is complete missing. . . . 24
18 screenshot for word count . . . 25

1 Introduction

1.1 Background & Motivation

Nowadays, the number of people with dementia is increasing because of the rising number of older people [9]. Furthermore, Wittenberg et al. [18] predicted the cost of dementia care in the UK in 2040 will be 172% cost in 2019. Therefore, people are looking for possible ways to prevent dementia. Livingston et al. [9] pointed out 12 potential risk factors (education, hypertension, hearing impairment, smoking, obesity, depression, physical inactivity, diabetes, low social contact, excessive alcohol consumption, traumatic brain injury (TBI), and air pollution) for dementia. Therefore, people could change their behavior based on the risk factors to decrease the probability of dementia. However, we do not know the specific relationship between the 12 risk factors and dementia. This report will use the Bayesian Network to explore the specific relationship and the direct risk factors. Additionally, this report analyzes the relationship between sex and dementia, Age and dementia, and the difference in the countries.

1.2 Data

SHARE provided the data collected from multiple countries at different times (8 waves from 2000 to 2020). By the way, the 8 waves mean the 8 surveys asked by 8 different periods. Moreover, the data include 7 types of variables (Demographics, Household composition, Social support & network, Childhood conditions, Health and health behavior, Functional limitation indices, and Work & money). Although, there are some risk factors not included in the data sets, such as DBI, air pollution, and hearing impairment. It contains the majority of risk factors, which we will study. This report will carry out two cross-sectional studies based on one wave & one country and multiple waves & countries.

2 Data pre-processing

2.1 Variable selection

At first, we need to extract the significant variables based on the 12 risk factors. Hearing impairment, TBI, and air pollution do not include in the data set, so we do not consider them in this report. For diabetes and hypertension, the data sets also do not consist of them, but I find a relative variable (chronic_mod) which records the number of diseases in the participant (10 types of diseases including hypertension and diabetes). For the other 7 risk factors, each risk factor can find at least one relative variable in the data set. Besides, poverty may also be related to dementia [9], so I choose a variable that records the household makes ends meet. Moreover, Beam et al. [1] mentioned that sex is related to dementia in the late 80s group people. Therefore, the sex variable was also considered in this report. The country has also consisted because Livingston et al. [9] pointed Mediterranean diet is good for dementia, and dementia in LMIC needs to investigate. In the end, we need a variable to classify people into three parts (Dementia, Cognitively impaired without dementia (CIND), and Normal). The cutoff point for dementia is determined by the mean of cognitive score minus 1.5 multiple standard deviation [16]. For the CIND, the boundary set by satisfying that around 20% people are CIND in the 70+ group because Crimmins et al. [3] mentioned that around 22% people are CIND in the US for people older than 70. There are 5 variables (recall_1,recall_2, orienti, numeracy_1, and numeracy_2) to measure the cognitive score using different ways in the data set. However, orienti, numeracy_1, and numeracy_2 are missing in some waves completely. Therefore, we only combine recall_1 and recall_2 for multiple wave study and use all 5 variables for the single wave study.

2.2 Data manipulation

After extracting the important variables, we also need to transform the variables to be suitable for the following research. For the structural learning in section 3, continuous variables can not be parents (X_j is parent of X_i if there is a directed edge from X_j to X_i) of category variables by *bnlearn* [13]. Therefore, I change all continuous variables to category variables to solve this problem. I also combine the

uncommon levels into another new level to avoid less information appearing. For example, few people have more than 3 diseases, so I put these people into the $[3, 11)$ category. The summary of all variables will illustrate as a table in the following, and you can find the detail of each variable in the appendix B.

variable's name (abbreviation)	origin name	category
sex (Sex)	female	0: male, 1: female
age (Age)	age	order factor; divide into $[60, 65), [65, 70), [70, 75), [75, 80), [80, 85), [85, 90), [90, 120)$
live_with_number_people (LWNP)	hhsize	order factor: divide into $(0, 1], (1, 2], (2, 13]$ This variable is using to measure social contact
depression_scale (DS)	eurod	order factor: divide into $[0, 2), [2, 5), [5, 13)$ higher value means more depressed
number_of_diseases (NOD)	chronic_mod	order factor: divide into $[0, 1), [1, 2), [2, 3), [3, 11)$
BMI (BMI)	bmi2	order factor: divide into 1: underweight, 2: normal, 3: overweight, 4: obese
Smoking (SMK)	smoking	factor variable: N (not smoking) Y (smoking)
Drinking_behavior (DK)	br010_mod	order factor: divide into $[1, 3), [3, 5), [5, 7), [7, 8)$ higher value means drinking more often
vigorous_activities (VA)	br015_	factor variable: divide into $[1, 4)$ means do some activities, $[4, 5)$ means hardly ever
Household_able_to_make_ends_meet (HAM)	co007_	order factor: divide into $1, 2, 3, 4$. higher value means easier
education_level_new (EDUL)	isced1997_r	order factor: divide into $[0, 2), [2, 3), [3, 5), [5, 7)$ higher value means higher education level
dementia_recall (AD)	combine_recall	This variable only use in the multiple wave study. order factor: divide into $(-1, 2]$: dementia, $(2, 5]$: CIND, $(5, 20]$: normal.
dementia_cognitive (AD)	combine_cognitive	This variable only use in the wave 4 study. order factor: divide into $(-1, 9]$: dementia, $(9, 15]$: CIND, $(15, 33]$: normal.
countries (C)	country_mod	This variable only use in the multiple countries study. factor variable: divide into 0: means developing countries, 1: means developed countries, 2: means developed countries & in Mediterranean.

Table 1: Variables information

2.3 Multiple Imputation

This report contains two studies. We first focus on one wave (wave 4) and one country (Estonia), then focuses on multiple waves (all waves except wave 3) and multiple countries (France, Switzerland, Italy, Spain, Bulgaria, and Croatia). For the first study, I choose to wave 4 and Estonia because wave 4 contains the fewer missing values base on the missing pattern for each wave in appendix C, and Estonia contains most people in wave 4. For multiple waves and countries, I combine the data from 7 waves except wave 3 because wave 3 has many variables missing completely. Moreover, I only keep the first occurrence for people who occurred multiple times in the data sets to avoid redundancy. I chose these countries because I want to explore the difference between the developing countries and developed countries, and the difference between Mediterranean countries and the central European countries.

After that, we need to deal with the missing values because of the massive of missing data. I used multiple imputations that impute the missing values by many times. The reason is that it can get less biased results [10]. For the imputations methods (using other observed variables to predict the missing variable base on regression), I chose *polr* regression for multiple categories variables, logistic regression for binary variables, and predictive mean matching for continuous variables. The continuous variable is the cognitive score (combine_recall or combine_cognitive) used to generate the dementia variables (dementia_recall or dementia_cognitive). All the missing variables are predicted based on all the other observed variables except the dementia variable to avoid multi-collinearity because it already contains the cognitive score variable as one of the predictors.

To measure the performance of the multiple imputations, I compare the distribution of imputed values and observed for each variable in the figure 1 and figure 2, density plots for continuous variables, and histograms for categorical variables. The density and histogram plots seem fine for all variables. However, I find more people classified as dementia or CIND than the observed value, and the density plots for missing values are on the left of the blue curve. These phenomenons mean the people who contain the missing value in the AD variable tend to have a higher risk. It may be because the people with the missing values on the AD variable are older. The following figure 3 does show this. Therefore, these multiple imputations are reasonable.

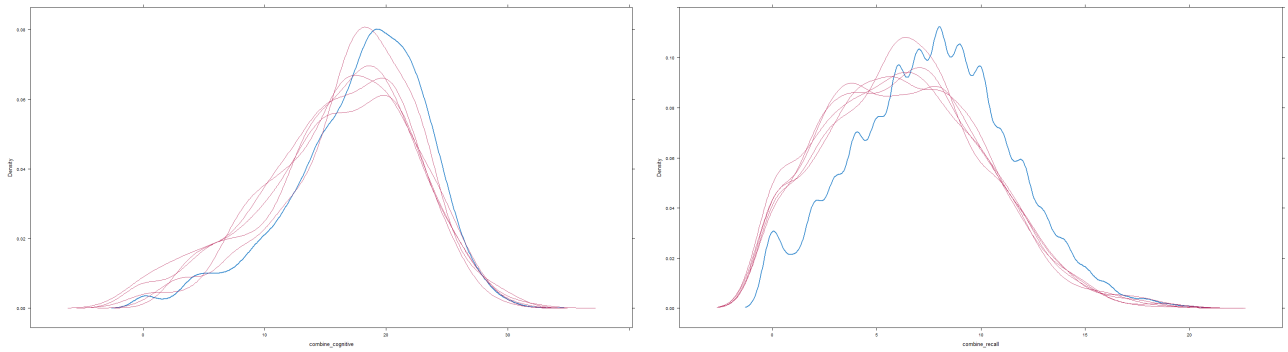

Figure 1: Compare observed values and missing values for the cognitive score. The left one is for wave 4 study. The right one is for multiple wave study. By the way, blue curve is the observed values, red curves are 5 different imputations' plots.

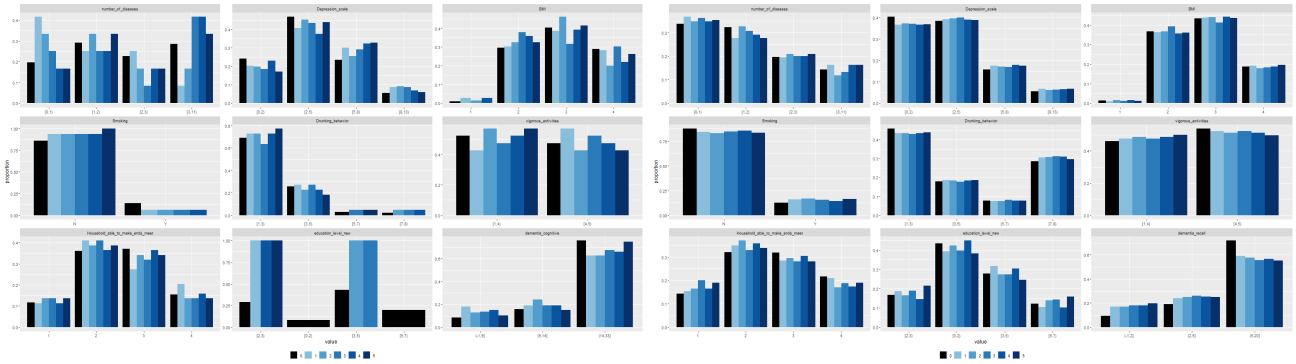

Figure 2: Compare observed values and missing values for categorical variables. The left one is for all missing category variables in the wave 4 study, the right one is for all missing category variables in the multiple wave study. The variables are NOD, DS, BMI, SMK, DK, VA, HAM, EDUL, and AD (from left to right and top to bottom). By the way, the black plots are the observed plot and the blue plots are 5 different imputations' plots.

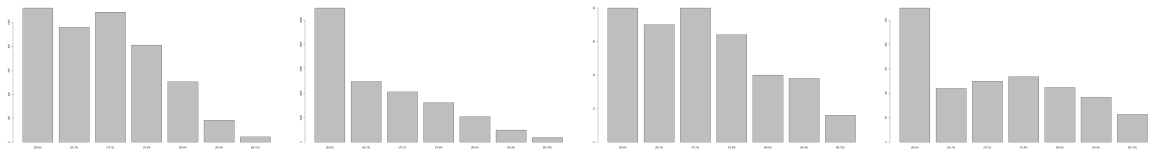

Figure 3: Distribution of age. Two plot on the left are observed distribution of age, the right are missing distribution of age.

3 Methods for Bayesian Network

Bayesian Network is a graphical model which is a directed acyclic graph (DAG/dag) containing nodes (random variables), edge with an arrow (represent the dependencies of each node) [7]. We can get the factorization of the joint probability distribution of all variables $(X_1, X_2, ..., X_n)$ based on the dag by the Markov property of Bayesian networks [7], the form is in the following:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|\Pi_{X_i}) \qquad (for\ discrete\ variables) \tag{3.1}$$

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} f(X_i|\Pi_{X_i}) \qquad (for\ continuous\ variables) \tag{3.2}$$

where Π_{X_i} means the parents of node X_i. Furthermore, the probability distribution of one node is only dependent on its parents. Therefore, there are two main steps in Bayesian network learning. The first step is learning the graphical structure, also called structural learning. After the structural learning step, the main thing is fitting the model base on the dag. The details of implementing them are in the next section 4.

Eventually, I want to point out a definition called Markov Blanket for DAG since it is significant for our analysis: Markov Blanket of a node X contains its parents, children (X_j is the child of X_i if there is a directed edge from X_i to X_j), and co-parents (the other parents of the children), denoted by MB(X) [6]. Furthermore, it has a property for MB(X):

$$X \perp\!\!\!\perp \{\ all\ variables \setminus X \setminus MB(X) \mid MB(X)\} \tag{3.3}$$

which means X is independent of other variables if we know the Markov Blanket of X.

4 Bayesian Network workflow

Firstly, I used different methods and data sets to learn the dag. Secondly, I used three types of network scores to capture the candidates. After getting the candidates, we need to compare them using the loss value of the k-fold validation to choose the best one. However, structural learning may not go through all possibilities, so I use the independence test to test whether it needs to add new edges or not. Nevertheless, we can not ensure whether the added edges do enhance the performance of the dag or not. Therefore, I test these edges by the loss value. I chose the dag with the lowest loss values. After that, we need to fit the model. I also used three types of data sets to fit the model and compare the performance by prediction error on the external data sets called external validation, which is necessary for the prediction model [2]. After we get the estimates, the probability distribution of AD will present as a line plot. However, If the figure plot is not smooth, it can be solved by refitting the model by *polr* regression. By the way, we also need to retest the performance of the new estimates. Eventually, I will present two diagnosis plots (calibration plot and ROC cure) to measure the performance.

4.1 Structural learning part

This report contains two types of algorithms which are constraint_based algorithms (gs,iamb,fast.iamb and inter.iamb) and score_based algorithms (Hill-Climbing with either bic or bde score) to learn the dag base on three types of data sets (data set with NA, data set drop NA and multiple imputations). However, the score_based algorithms can not allow NA in the data sets in *bnlearn* [13]. Therefore, constraint-based algorithms only utilized the NA data set and the 5 imputations, and score-based algorithms only used the drop NA data set and the imputations.

For the detail of the structural learning, the first step is setting the blacklist that the edges we do not want to appear in the dag, and the whitelist that we certainly should appear in the dag.

Secondly, I apply the bootstrap techniques to the structural learning to get more robust results [10]. The basic idea is re-sampling the data set multiple times and doing the structural learning many times, then keeping the edges which often appear in the plenty of dags [4].

After that, compare the different graphical structures learned from different algorithms with different types of data sets by three different network scores (Bayesian Information Criterion (bic) score, Bayesian Dirichlet equivalent (bde) score, and log-likelihood score). It may get three candidates because of three different scores. I decide to apply k-fold cross-validation (k = 10) [11] which splits the data sets into K parts and uses k-1 parts to fit the model, and 1 part to do cross-validation. After that, I choose the dag with the lowest loss value.

Eventually, structural learning can not consider all probability due to the massive combination of the edges. Therefore, we can use the independent test (Jonckheere-Terpstra (jt) test for ordered variables, and mutual information (mi) test for others)to test the dependence of two nodes either conditional or unconditional. If it does illustrate that dementia dependent on some nodes under dementia's Markov Blanket, we could add the edges from these nodes direct to the dementia node. However, the new graph with the added edges needs to be tested by loss value.

4.2 Fit model part

After getting the graphical structure, we need to fit the model (learn the joint probability distribution) based on the structure. As I mentioned before, we can use three types of data sets (with NA, drop NA, and multiple imputations) to fit the model. To avoid overfitting, I used k-fold cross-validation (k = 10) [11]. I choose the estimate with the lowest loss value of the 10 different estimates. However, the data set with NA can not apply the k-fold cross-validation, so we do not apply this technique to the data set with the NA. Moreover, It is better to use new data sets to compare the performance of a predicted model with different estimates called external validation. Therefore, I used the data of wave 5 for the wave 4 model and another three countries' data for the multiple countries model to compare the different estimates by prediction error. Eventually, I chose the estimates with the lowest predictor error.

4.3 Present the estimates

I use a par-plot to show how the parents of dementia (AD) be affected and a line plot to demonstrate how the parents of dementia affect it. However, the line plot may not be smooth. It can solve by using *plor* regression to refit the joint probability distribution of dementia with its parents.

4.4 Diagnosis

I use two ways to measure the performance of the models, which are the calibration plot and the ROC curve to test the discrimination [15]. Moreover, I also calculate the area under the curve (AUC) to measure the discrimination.

5 Results

I will separate this into two parts. One is the results of wave 4 in Estonia, the other is the results of multiple waves and multiple countries. By the way, I will present the results base on the workflow.

5.1 wave 4 & Estonia

For the blacklist, I set all edges direct to Sex or Age and all edges from AD (dementia) to others. However, I exclude the edge from Sex to Age. The reason is that females tend to live longer than males. For the whitelist, I only set only one edge from EDUL (education level) to AD (dementia)

based on the article written by Livingston et al. [9].

After that, I use three types of scores to compare different graphs learned from different data sets by different methods. The three tables 2 3 4 are in the following. We can see that graph 18 (row 18) has the best performance on bic and bde score overall, and graph 8 (row 8) has the best performance on loglik score. Therefore, we get two graphical structures, graph 8 and graph 17.

To compare these two graphs, compare their loss values from k-fold cross-validation. A lower loss value means better, so graph 18 is the better one from the table 5.

		1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
gs	with NA	-56192.29	-56159.83	-56168.48	-56174.87	-56119.09	-50363.16
gs	1^{st} imputation	-60678.97	-60647.01	-60656.57	-60674.05	-60614.62	-54756.09
gs	2^{ed} imputation	-60678.97	-60647.01	-60656.57	-60674.05	-60614.62	-54756.09
iamb	with NA	-56412.81	-56387.71	-56408.89	-56408.52	-56357.41	-50689.62
iamb	1^{st} imputation	-56470.83	-56438.45	-56463.64	-56460.86	-56401.74	-50729.16
iamb	2^{ed} imputation	-56641.92	-56609.67	-56632.83	-56629.77	-56570.99	-50897.69
fast.iamb	with NA	-56672.97	-56641.04	-56660.75	-56660.71	-56607.60	-50928.25
fast.iamb	1^{st} imputation	-57490.01	-57463.77	-57484.76	-57480.17	-57440.31	-51779.24
fast.iamb	2^{ed} imputation	-57490.01	-57463.77	-57484.76	-57480.17	-57440.31	-51779.24
inter.iamb	with NA	-56412.81	-56387.71	-56408.89	-56408.52	-56357.41	-50689.62
inter.iamb	1^{st} imputation	-56470.83	-56438.45	-56463.64	-56460.86	-56401.74	-50729.16
inter.iamb	2^{ed} imputation	-56641.92	-56609.67	-56632.83	-56629.77	-56570.99	-50897.69
hc.bic	drop NA	-55545.73	-55525.02	-55535.31	-55546.48	-55480.78	-49786.54
hc.bic	1^{st} imputation	-55462.20	-55442.43	-55449.20	-55465.38	-55397.30	-49714.53
hc.bic	2^{ed} imputation	-55444.70	-55422.67	-55433.07	-55447.22	-55379.31	-49699.67
hc.bde	drop NA	-55531.09	-55515.75	-55520.20	-55533.86	-55470.12	-49760.42
hc.bde	1^{st} imputation	-55439.97	-55420.11	-55427.53	-55445.13	-55381.78	-49711.58
hc.bde	2^{ed} imputation	-55425.33	-55402.95	-55413.25	-55428.70	-55364.06	-49704.35

Table 2: bic network score for different graphical structure from different algorithm with different data sets (wave 4 Estonia)

		1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
gs	with NA	-55991.15	-55956.36	-55963.44	-55970.34	-55913.55	-50153.42
gs	1^{st} imputation	-58874.04	-58811.17	-58836.16	-58846.53	-58786.93	-52805.86
gs	2^{ed} imputation	-58874.04	-58811.17	-58836.16	-58846.53	-58786.93	-52805.86
iamb	with NA	-55890.40	-55858.56	-55883.02	-55891.14	-55823.55	-50110.06
iamb	1^{st} imputation	-56029.92	-55990.59	-56019.51	-56024.73	-55949.46	-50231.54
iamb	2^{ed} imputation	-56140.19	-56100.77	-56127.72	-56132.57	-56057.69	-50336.89
fast.iamb	with NA	-56138.88	-56100.19	-56123.26	-56131.59	-56062.13	-50336.66
fast.iamb	1^{st} imputation	-56476.57	-56444.61	-56480.50	-56474.46	-56414.68	-50682.13
fast.iamb	2^{ed} imputation	-56476.57	-56444.61	-56480.50	-56474.46	-56414.68	-50682.13
inter.iamb	with NA	-55890.40	-55858.56	-55883.02	-55891.14	-55823.55	-50110.06
inter.iamb	1^{st} imputation	-56029.92	-55990.59	-56019.51	-56024.73	-55949.46	-50231.54
inter.iamb	2^{ed} imputation	-56140.19	-56100.77	-56127.72	-56132.57	-56057.69	-50336.89
hc.bic	drop NA	-55354.22	-55332.19	-55342.63	-55354.05	-55287.84	-49591.40
hc.bic	1^{st} imputation	-55320.57	-55300.69	-55307.50	-55323.70	-55255.54	-49571.24
hc.bic	2^{ed} imputation	-55296.15	-55273.97	-55284.42	-55298.59	-55230.58	-49549.35
hc.bde	drop NA	-55398.63	-55383.30	-55387.74	-55401.40	-55337.67	-49625.89
hc.bde	1^{st} imputation	-55244.05	-55224.22	-55231.70	-55249.35	-55185.98	-49512.00
hc.bde	2^{ed} imputation	-55210.25	-55187.73	-55198.14	-55213.68	-55148.89	-49485.34

Table 3: bde network score for different graphical structure from different algorithm with different data sets (wave 4 Estonia)

		1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
gs	with NA	-55325.53	-55293.07	-55301.72	-55308.11	-55252.33	-49507.02
gs	1^{st} imputation	-55089.44	-55057.48	-55067.04	-55084.53	-55025.09	-49235.05
gs	2^{ed} imputation	-55089.44	-55057.48	-55067.04	-55084.53	-55025.09	-49235.05
iamb	with NA	-54518.62	-54493.53	-54514.71	-54514.34	-54463.23	-48818.65
iamb	1^{st} imputation	-54716.17	-54683.80	-54708.99	-54706.21	-54647.09	-48996.01
iamb	2^{ed} imputation	-54620.89	-54588.65	-54611.81	-54608.74	-54549.97	-48901.43
fast.iamb	with NA	-54715.37	-54683.44	-54703.15	-54703.10	-54650.00	-48994.63
fast.iamb	1^{st} imputation	-54357.00	-54330.76	-54351.75	-54347.16	-54307.30	-48684.62
fast.iamb	2^{ed} imputation	-54357.00	-54330.76	-54351.75	-54347.16	-54307.30	-48684.62
inter.iamb	with NA	-54518.62	-54493.53	-54514.71	-54514.34	-54463.23	-48818.65
inter.iamb	1^{st} imputation	-54716.17	-54683.80	-54708.99	-54706.21	-54647.09	-48996.01
inter.iamb	2^{ed} imputation	-54620.89	-54588.65	-54611.81	-54608.74	-54549.97	-48901.43
hc.bic	drop NA	-54755.08	-54734.36	-54744.66	-54755.83	-54690.13	-49005.58
hc.bic	1^{st} imputation	-54866.04	-54846.27	-54853.04	-54869.22	-54801.14	-49125.68
hc.bic	2^{ed} imputation	-54823.18	-54801.14	-54811.55	-54825.69	-54757.78	-49085.76
hc.bde	drop NA	-54968.75	-54953.41	-54957.86	-54971.52	-54907.78	-49204.97
hc.bde	1^{st} imputation	-54742.34	-54722.48	-54729.90	-54747.50	-54684.15	-49022.50
hc.bde	2^{ed} imputation	-54651.59	-54629.21	-54639.51	-54654.96	-54590.32	-48940.09

Table 4: log likelihood network score for different graphical structure from different algorithm with different data sets (wave 4 Estonia)

graph	1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
graph 8	11.64	11.60	11.65	11.67	11.61	11.55
graph 17	11.49	11.57	11.47	11.47	11.49	11.51

Table 5: compare two graph by bn.cv (wave 4 Estonia)

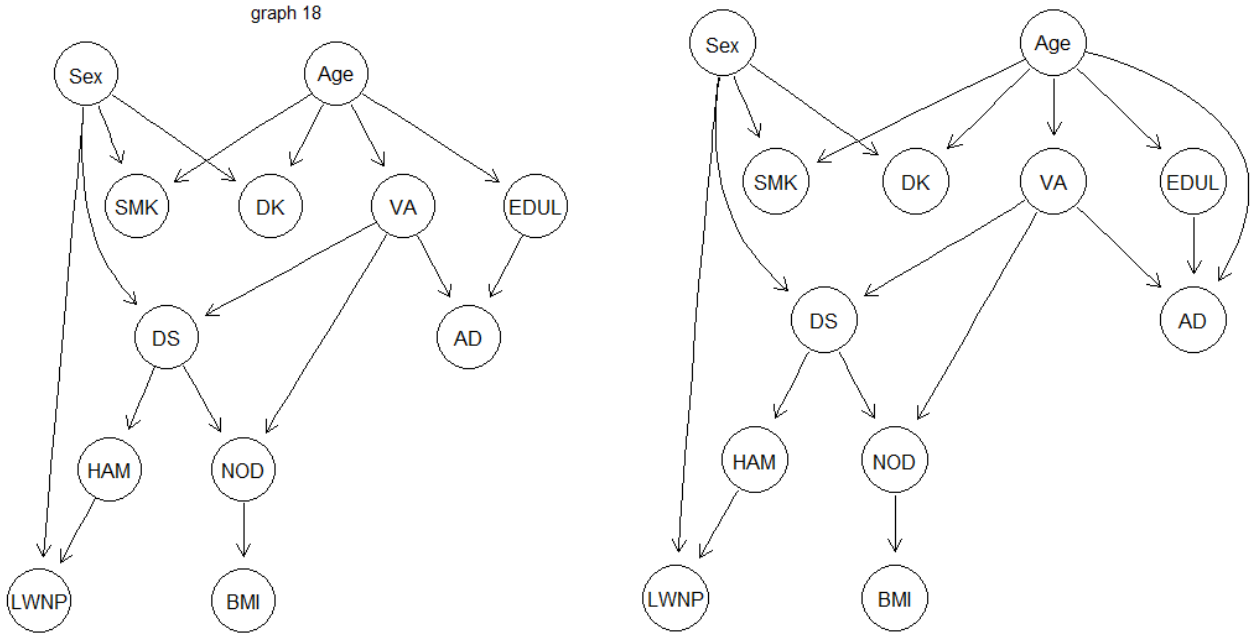

Figure 4: left is dag 18 & right is better dag

After we learned the best graphical structure (graph 18) in the left part of the figure 4, we can use the independence test to add some edges, as I mentioned before in the section 4.1. Based on the Markov Blanket property 3.3, AD (dementia) is independent of all other nodes under the VA (vigorous activities) and EDUL (education level) conditions. I find the independence test show dependence on AD with Age (p-value $<2.2 \cdot 10^{-16}$), DS (p-value $<2.2 \cdot 10^{-16}$) and BMI (p-value <0.001). However, whether adding these three edges strengthen the performance is a question. Therefore, I used the loos value to test the performance of the new graph after adding three edges in the table 6. Based on the table 6, we only need to add the edge from Age to AD, and the final dag is in the right of the figure 4.

graph	1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
graph 18	11.52	11.53	11.57	11.52	11.56	11.44
add Age	11.48	11.57	11.45	11.46	11.48	11.51
add DS	11.52	11.45	11.57	11.53	11.46	11.47
add BMI	11.52	11.58	11.61	11.55	11.51	11.45
add Age & DS	11.58	11.52	11.57	11.54	11.52	11.53
add BMI & DS	11.63	11.48	11.46	11.57	11.59	11.53
add Age & BMI	11.63	11.56	11.53	11.50	11.52	11.51
add all three	11.78	11.81	11.76	11.70	11.80	11.73

Table 6: compare graphs add edges by bn.cv (wave 4 & Estonia)

After getting the final dag, we can fit the model base on the dag. I apply the k-fold (k=10) validation to avoid overfitting, as I mentioned before in the section 4.2. By the way, I also used three types of data sets to fit the model. For multiple imputations, Ii will get 5 different estimates because of 5 times imputations. However, I take the average of the 5 estimates according to Rubin's Rule [12]. Therefore, there are three estimates. One is the estimate from the data with NA, the others from the drop NA data sets, and the multiple imputations. To compare these three estimates, I used the external validation base on the new data sets (the wave 5 & Estonia's data) to see the prediction error. The estimates from the imputations and data sets with NA both get the lowest prediction

error is 0.2178218, and the other prediction error is 0.2326733 (drop NA). Therefore, we choose the estimates based on the data sets with NA because imputations may bias. However, It does illustrate that the multiple imputations are similar to the original data sets.

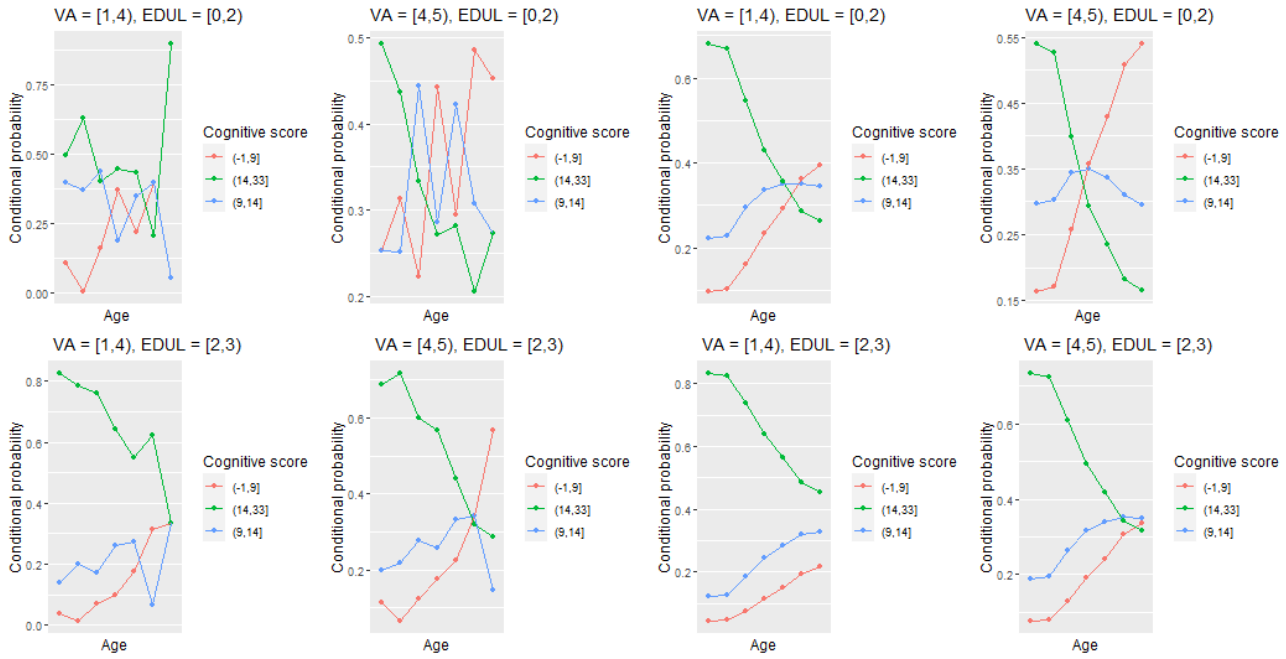

Figure 5: relation of AD

From the dag, we can see that AD (dementia) can be predicted, if we only know VA, Age, and EDUL. Therefore, the direct factors are VA, Age, and EDUL based on the dag. However, the plot is not smooth in the left part of figure 5. I used *polr* regression to solve it. To see the performance of the new estimates, we need to recalculate the prediction error as 0.2277228, which is reasonable. The new plot is the right part of figure 5 which could illustrate how these three nodes affect dementia. We can get older people to have a higher risk for dementia from the new plots. Comparing plot 1 (Top left) with the other 3 plots on the right of figure 5, we can see people who do activities get a lower risk, and people who have a higher level of education experience get a lower risk.

For the parents of the parents of AD are demonstrated by par-plot in the figure 6. For VA: Age point to VA, we can see older people do fewer activities from the figure 6. For EDUL: Age point to EDUL, we can see the elderly have lower education levels from the figure 6. For Age: it does not have parents.

For diagnosis of the estimates, I use two ways which are calibration plot and the ROC curve to measure the discrimination. For the calibration plot, we can see all the points are near the line, and the confidence intervals of these points pass through the line. Therefore, it is a valuable overall and local calibration plot based on the article written by Stevens et al. [14]. For the ROC curve, it is beyond the $y = x$ line, which means the predicted model is better than the random pick according to Hanjian [5]. Furthermore, the area under the curve (AUC) is 0.8074 approach to 1, which seems reasonable. The diagnoses seem fine for the estimate, so this model is good overall.

We already know Age affects dementia, so I would like to talk about the relationship between Sex and dementia. For a directed graph, we can use the global Markov property [8] to determine the independence of two nodes. Sex is independent of AD (dementia) based on the property because all the trails from Sex to AD contain at least one collider ($A \to B \leftarrow C$). Therefore, Sex is not related to the dementia base on this model.

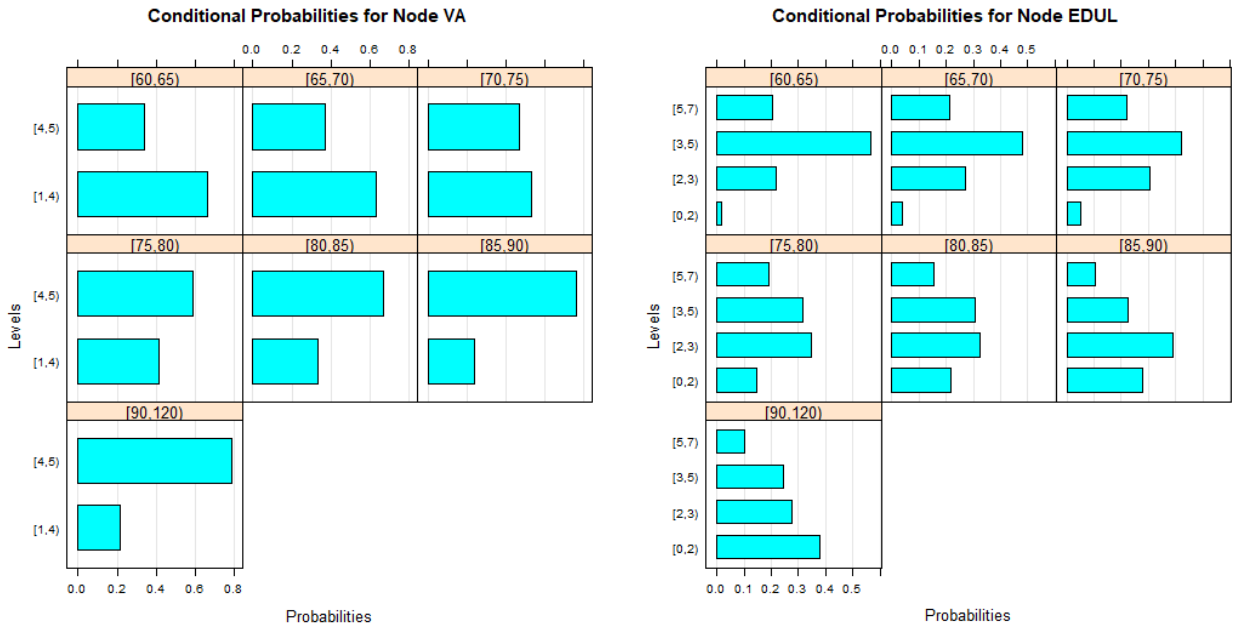

Figure 6: Markov blanket variables of AD

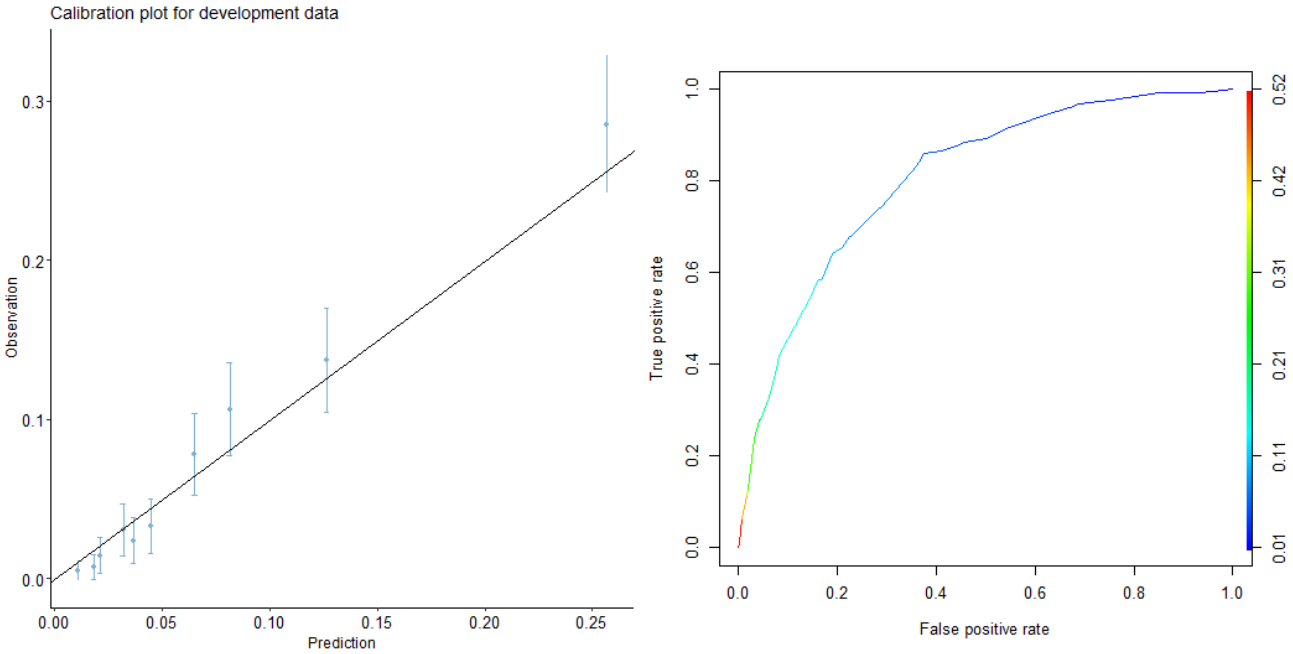

Figure 7: diagnosis_1

5.2 multiple waves & multiple countries

For the blacklist, I set all edges direct to Sex, Age, or C (country) and all edges from AD (dementia) to others. However, I exclude the edge from Sex to Age, country to Sex, and country to Age. The reason is that females tend to live longer than males, and different countries have different demographic compositions. The whitelist is as same as before.

After that, I use three types of scores to compare different graphs based on different data sets. The three tables 7 8 9 are in the following. We can see that graph 17 (row 17) has the best performance on bic, graph 14 (row 14) has the best performance on bde, and graph 8 (row 8) has the best performance on loglik score. Therefore, we get three graphical structures graph 8, graph 14, and graph 17.

		1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
gs	with NA	-354863.92	-354871.01	-354852.97	-354722.66	-355031.23	-218253.31
gs	1^{st} imputation	-368750.84	-368795.14	-368897.15	-368703.61	-368964.80	-230969.14
gs	2^{ed} imputation	-365841.14	-365933.69	-366034.48	-365784.94	-366095.07	-228253.47
iamb	with NA	-345930.58	-345960.54	-346038.14	-345925.98	-346159.88	-213442.87
iamb	1^{st} imputation	-347461.74	-347428.98	-347566.07	-347410.12	-347636.54	-214762.26
iamb	2^{ed} imputation	-349090.37	-349050.72	-349206.49	-349062.96	-349253.27	-216383.14
fast.iamb	with NA	-354573.36	-354618.13	-354722.28	-354575.65	-354800.48	-218831.38
fast.iamb	1^{st} imputation	-352339.43	-352340.32	-352398.37	-352323.40	-352520.47	-219577.51
fast.iamb	2^{ed} imputation	-350705.10	-350701.43	-350778.71	-350666.09	-350885.29	-218050.18
inter.iamb	with NA	-345930.58	-345960.54	-346038.14	-345925.98	-346159.88	-213442.87
inter.iamb	1^{st} imputation	-347461.74	-347428.98	-347566.07	-347410.12	-347636.54	-214762.26
inter.iamb	2^{ed} imputation	-349090.37	-349050.72	-349206.49	-349062.96	-349253.27	-216383.14
hc.bic	drop NA	-343710.70	-343729.02	-343823.27	-343698.97	-343878.79	-210962.70
hc.bic	1^{st} imputation	-343493.33	-343513.61	-343624.08	-343483.85	-343677.91	-211234.03
hc.bic	2^{ed} imputation	-343493.33	-343513.61	-343624.08	-343483.85	-343677.91	-211234.03
hc.bde	drop NA	-344066.02	-344075.10	-344158.97	-344040.81	-344231.70	-211143.79
hc.bde	1^{st} imputation	-343494.35	-343504.93	-343622.18	-343481.61	-343665.73	-211086.56
hc.bde	2^{ed} imputation	-343494.35	-343504.93	-343622.18	-343481.61	-343665.73	-211086.56

Table 7: bic network score for different graphical structure from different algorithm with different data sets (multiple waves & multiple countries)

To compare these three graphs, compare their loss values, get from k-fold cross-validation. A lower loss value means better, so graph 14 is the better one from the table 10.

		1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
gs	with NA	-354469.36	-354466.08	-354458.90	-354327.04	-354634.76	-217799.72
gs	1^{st} imputation	-364654.14	-364714.08	-364762.14	-364618.59	-364832.72	-226024.56
gs	2^{ed} imputation	-362648.94	-362766.02	-362851.53	-362584.97	-362892.58	-224386.31
iamb	with NA	-345176.01	-345201.84	-345284.64	-345174.12	-345414.00	-212097.55
iamb	1^{st} imputation	-346496.94	-346464.61	-346612.45	-346442.82	-346678.90	-213050.32
iamb	2^{ed} imputation	-347778.35	-347713.88	-347876.04	-347715.31	-347920.52	-214156.94
fast.iamb	with NA	-352905.05	-352909.71	-353051.01	-352891.27	-353119.71	-216455.87
fast.iamb	1^{st} imputation	-350003.23	-350001.22	-350027.18	-350017.36	-350191.66	-215592.23
fast.iamb	2^{ed} imputation	-348767.40	-348751.31	-348824.79	-348753.74	-348984.96	-214532.14
inter.iamb	with NA	-345176.01	-345201.84	-345284.64	-345174.12	-345414.00	-212097.55
inter.iamb	1^{st} imputation	-346496.94	-346464.61	-346612.45	-346442.82	-346678.90	-213050.32
inter.iamb	2^{ed} imputation	-347778.35	-347713.88	-347876.04	-347715.31	-347920.52	-214156.94
hc.bic	drop NA	-343299.94	-343318.30	-343412.68	-343288.03	-343468.33	-210323.83
hc.bic	1^{st} imputation	-342982.34	-343002.57	-343113.28	-342973.07	-343167.79	-210398.86
hc.bic	2^{ed} imputation	-342982.34	-343002.57	-343113.28	-342973.07	-343167.79	-210398.86
hc.bde	drop NA	-343678.31	-343687.33	-343771.41	-343652.81	-343844.21	-210528.38
hc.bde	1^{st} imputation	-343022.92	-343033.57	-343150.99	-343009.80	-343194.79	-210368.33
hc.bde	2^{ed} imputation	-343022.92	-343033.57	-343150.99	-343009.80	-343194.79	-210368.33

Table 8: bde network score for different graphical structure from different algorithm with different data sets (multiple waves & multiple countries)

		1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
gs	with NA	-351994.61	-352001.70	-351983.67	-351853.36	-352161.93	-215515.74
gs	1^{st} imputation	-351295.49	-351339.79	-351441.79	-351248.26	-351509.45	-214315.14
gs	2^{ed} imputation	-351168.46	-351261.00	-351361.79	-351112.25	-351422.39	-214254.40
iamb	with NA	-340859.61	-340889.56	-340967.17	-340855.01	-341088.91	-208604.70
iamb	1^{st} imputation	-341065.69	-341032.93	-341170.02	-341014.07	-341240.49	-208659.85
iamb	2^{ed} imputation	-340767.87	-340728.21	-340883.98	-340740.45	-340930.76	-208442.71
fast.iamb	with NA	-346678.95	-346723.72	-346827.88	-346681.24	-346906.07	-211299.40
fast.iamb	1^{st} imputation	-339980.53	-339981.42	-340039.47	-340161.57	-339964.50	-207786.00
fast.iamb	2^{ed} imputation	-340135.06	-340131.39	-340208.66	-340096.05	-340315.25	-207965.40
inter.iamb	with NA	-340859.61	-340889.56	-340967.17	-340855.01	-341088.91	-208604.70
inter.iamb	1^{st} imputation	-341065.69	-341032.93	-341170.02	-341014.07	-341240.49	-208659.85
inter.iamb	2^{ed} imputation	-340767.87	-340728.21	-340883.98	-340740.45	-340930.76	-208442.71
hc.bic	drop NA	-341682.31	-341700.63	-341794.88	-341670.58	-341850.40	-209027.43
hc.bic	1^{st} imputation	-340731.05	-340751.33	-340861.81	-340721.57	-340915.63	-208598.57
hc.bic	2^{ed} imputation	-340731.05	-340751.33	-340861.81	-340721.57	-340915.63	-208598.57
hc.bde	drop NA	-342129.37	-342138.45	-342222.31	-342104.15	-342295.05	-209296.05
hc.bde	1^{st} imputation	-341099.02	-341109.60	-341226.85	-341086.28	-341270.40	-208801.19
hc.bde	2^{ed} imputation	-341099.02	-341109.60	-341226.85	-341086.28	-341270.40	-208801.19

Table 9: log likelihood network score for different graphical structure from different algorithm with different data sets (multiple waves & multiple countries)

graph	1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
graph 8	12.75	12.79	12.77	12.81	12.84	12.51
graph 14	12.68	12.70	12.70	12.73	12.73	12.45
graph 17	12.68	12.72	12.71	12.74	12.75	12.47

Table 10: compare two graph by bn.cv (multiple waves & multiple countries)

After we learned the better graph (graph 14), we can use the independence test to add some edges as before. Based on the Markov Blanket property 3.3, AD (dementia) is independent of all other nodes under the Age (vigorous activities) and EDUL (education level) conditions. I find the dependence test show dependence on AD with VA (p-value $<2.2 \cdot 10^{-16}$), DS (p-value $<2.2 \cdot 10^{-16}$), HAM (p-value $<2.2 \cdot 10^{-16}$) and LWNP (p-value $<1 \cdot 10^{-5}$). However, whether the performance can be enhanced by adding these four edges or not. Therefore, I used loos value to test the performance of the new graph after adding four edges in the table 11. Base on the table 11, we only need add the edge from DS to AD, the final dag is illustrated in the right part of the figure 8.

graph	1^{st} imputation	2^{ed} imputation	5^{rd} imputation	4^{th} imputation	5^{th} imputation	drop_NA
graph 14	12.68	12.70	12.70	12.73	12.73	12.45
add VA	12.72	12.74	12.71	12.71	12.67	12.43
add DS	12.68	12.68	12.70	12.71	12.73	12.41
add HAM	12.73	12.74	12.72	12.73	12.68	12.42
add LWNP	12.69	12.73	12.74	12.71	12.73	12.41
add VA & DS	12.71	12.72	12.73	12.73	12.72	12.45
add VA & HAM	12.72	12.74	12.72	12.66	12.75	12.39
add VA & LWNP	12.75	12.71	12.74	12.69	12.74	12.42
add DS & HAM	12.75	12.72	12.76	12.72	12.76	12.41
add DS & LWNP	12.71	12.73	12.72	12.72	12.74	12.46
add LWNP & HAM	12.72	12.71	12.70	12.76	12.76	12.41
add VA & DS & HAM	12.76	12.76	12.80	12.79	12.78	12.52
add VA & DS & LWNP	12.73	12.79	12.75	12.74	12.75	12.44
add VA & LWNP & HAM	12.76	12.75	12.74	12.77	12.73	12.48
add LWNP & DS & HAM	12.74	12.73	12.71	12.71	12.72	12.46
add VA & DS & HAM &LWNP	12.77	12.76	12.73	12.74	12.76	12.45

Table 11: compare graphs add edges by bn.cv (multiple waves & multiple countries)

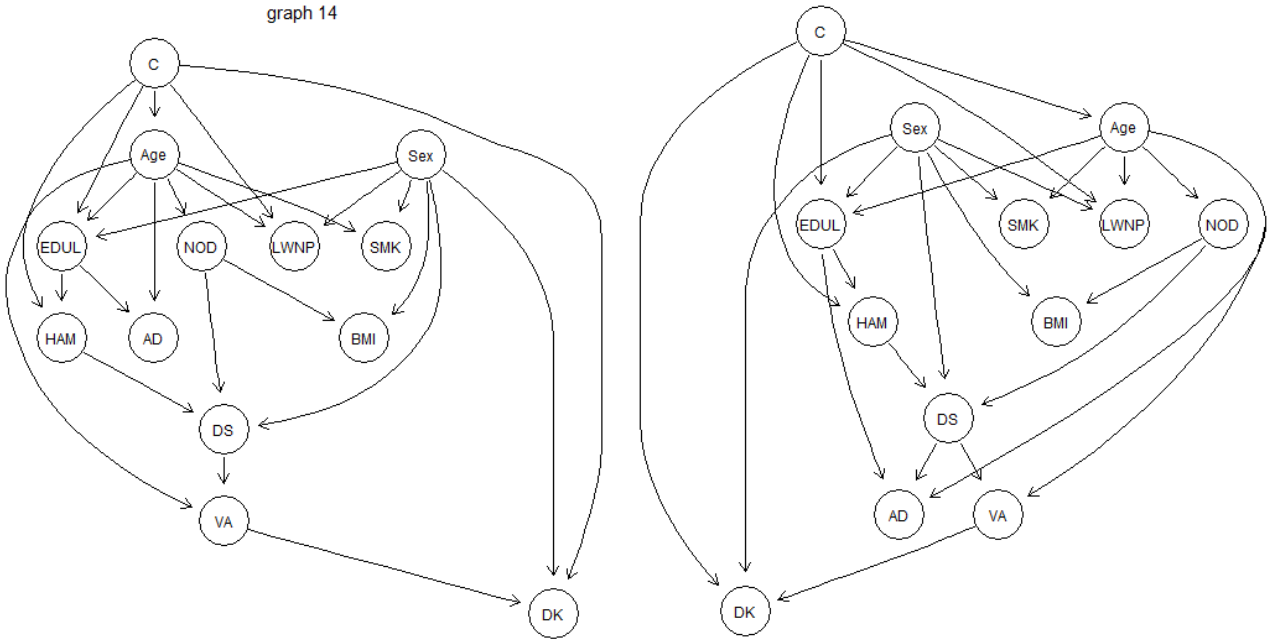

Figure 8: left is dag 14 & right is better dag

Same as before, I compare three estimates based on prediction error. The estimates learned from the drop NA data sets get the lowest prediction error is 0.1877386, and the others' prediction errors are 0.1887279 (imputations) and 0.1885746 (with NA). However, we can not choose the estimates based on the drop NA data set because the drop NA data set does not contain the level of the developing country. Therefore, we choose the estimates generated by data sets with NA.

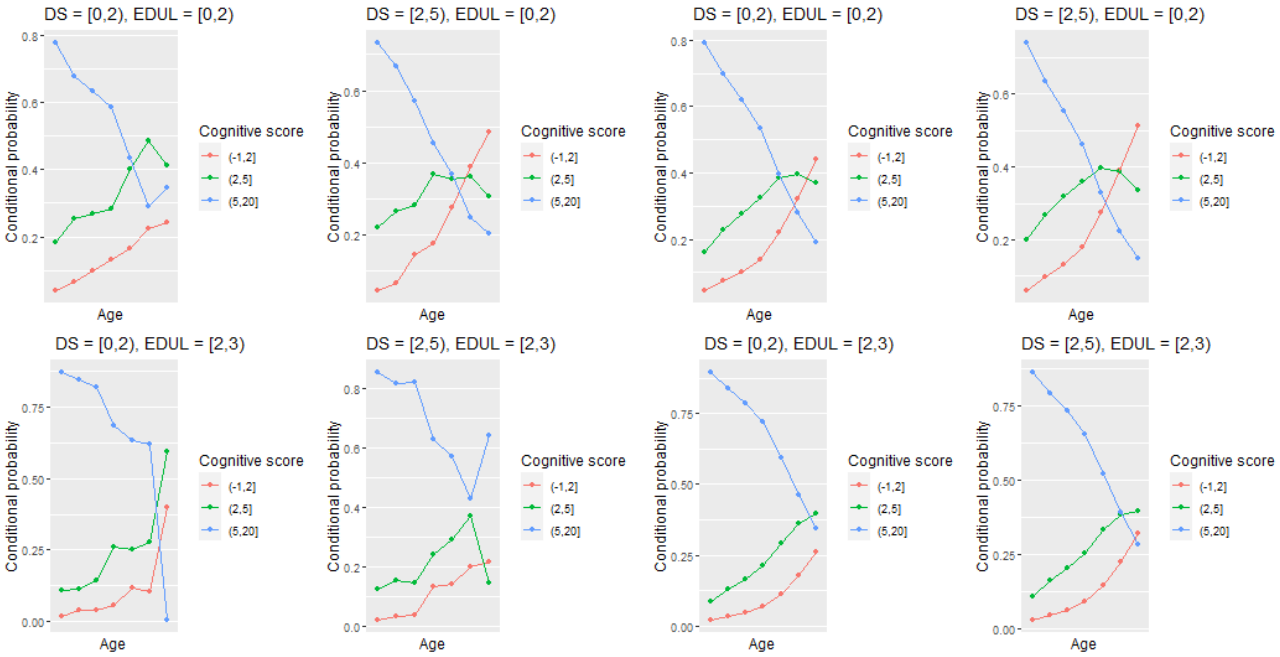

Figure 9: relation of AD

From the dag, we know AD (dementia) could be predicted, if we only know DS, Age, and EDUL. Therefore, the immediate risk factors are VA, Age, and EDUL based on the dag. However, the plot is not smooth in the left part of the figure 9. We use *polr* regression to solve it. To see the performance of the new estimates, we need to recalculate the predictor error, which is 0.1869862 better. The new plot is the right part of figure 5. From the new plot, we can get older people to have a higher risk for dementia. By comparing plot 1 (Top left) with the other 3 plots on the right of figure 9, we can see people are open, or people who have a higher level of education experience get a lower risk.

After that, I will discuss the nodes directed to DS, Age, and EDUL and their relationship. For DS: there are three nodes direct to DS, which are HAM, NOD and Sex. We can see females are easier be depressed than males, and people who get more diseases or have more difficulty paying their household expenditures, tend to be depressed easily from figure 10. For Age: Only country direct to Age.Different country has different age pattern from the figure 10. For EDUL: there are also three nodes (country, Age, and Sex) direct to EDUL. We can see the elderly have lower education level, the male gets higher education level, and different country has a different pattern from the figure 10.

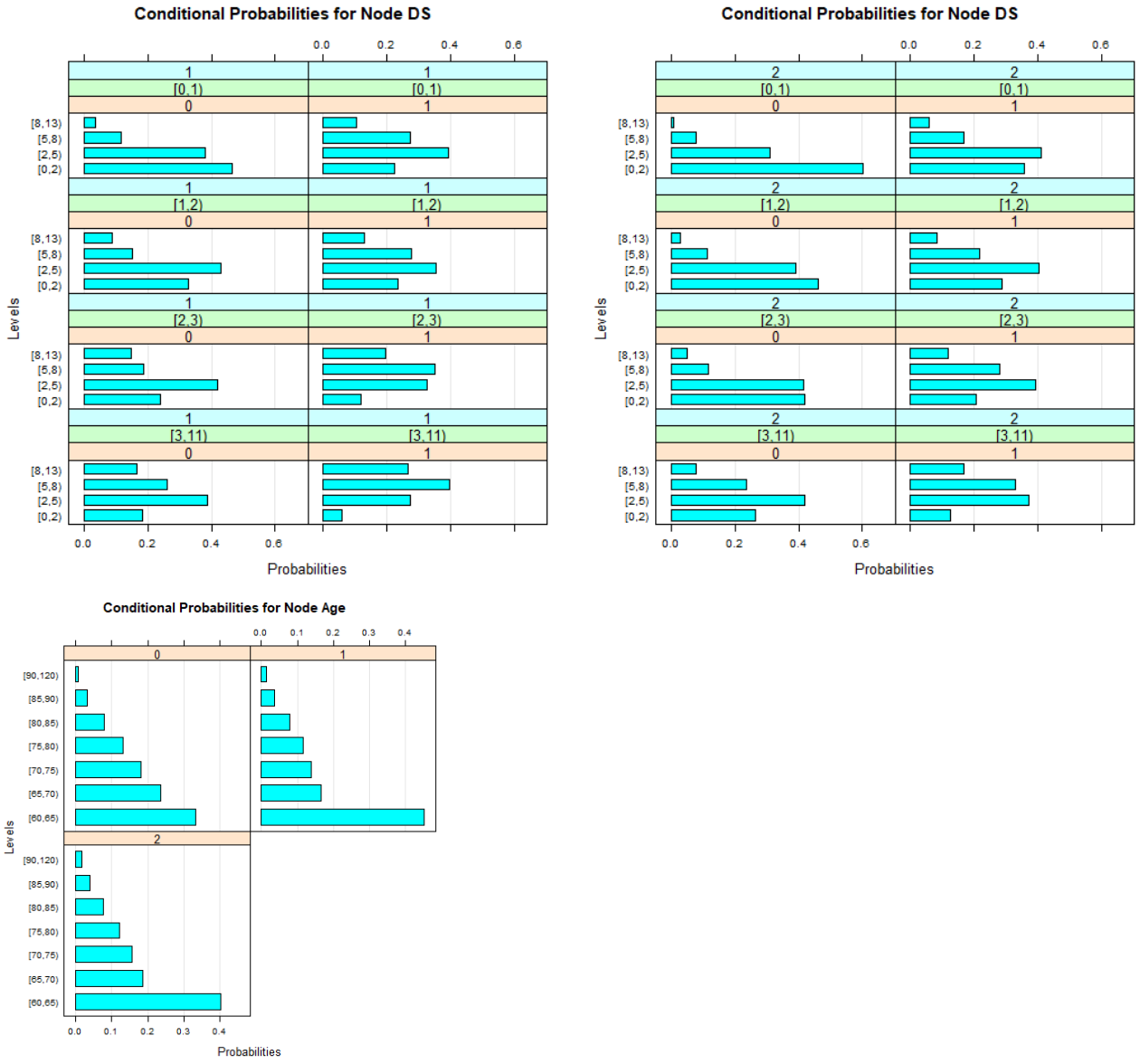

Figure 10: Markov blanket variables of AD_1. Variables for DS are HAM, NOD and Sex (Top to bottom). Variable for Age is country.

Figure 11: Markov blanket variables of AD_2. Variables for EDUL are country, Age, and Sex (Top to bottom)

For diagnosis, we can see all the points are near the line, and the confidence intervals of these points pass through the line excluded at the tail of the line in the calibration plot. Therefore, it is a valuable overall calibration plot base on the article written by Stevens et al. [14]. For the ROC curve, it is beyond the $y = x$ line, which means the predicted model is better than the random pick according to Hajian [5]. Furthermore, the area under the curve (AUC) is 0.8028, which is reasonable. The diagnoses seem fine for the estimate.

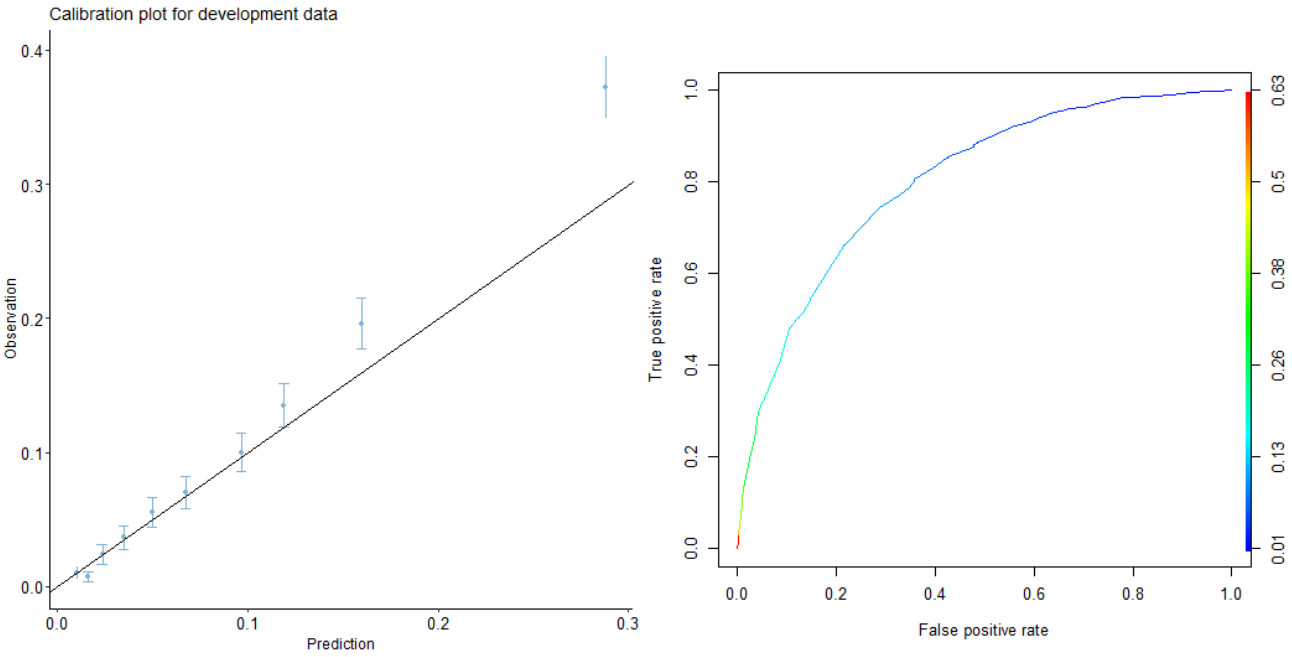

Figure 12: diagnosis_2

In the end, I also want to explore the relationship between Sex and AD (dementia), and country and AD in the multiple waves study. In this case, sex does related to the AD nodes base on the graph. Females tend to be more depressed and less educated than males, which both have a negative influence on dementia. For country node, it does influence dementia in different ways, such as EDUL (different education policy), Age (different demographic composition), and HAM (different Gini coefficient). However, dementia is independent of sex and country if we know the EDUL, Age, and DS by the Markov Blanket property 3.3. Therefore, dementia is indirectly dependent on sex and country nodes.

6 Conclusion

In this report, I use the Bayesian network to explore the potential direct risk factor of dementia, the relationship between Sex and dementia, and country and dementia. Both two models indicate age and education are parents of dementia. Therefore, these two variables affect dementia directly if no other unobserved con-founders. Moreover, activities and depression are connected, and one of them is direct to dementia in both models. Thus, activities and depression may also directly affect dementia under the no other unobserved con-founders assumptions. Therefore, people can prevent dementia by doing some activities, staying optimistic, and getting a higher education level. However, it does exist some risk factors exclude from the model. Thus, we can not ensure these nodes are still direct to dementia if we add new risk factors such as hearing impairment, TBI, and air pollution. Although we could not ensure which risk factors directly affect dementia, we do guarantee dementia independent of sex and countries under its Markov blanket. Therefore, dementia is not directly related to sex and country. However, there are several limitations in this report. For example, it does not include all 12 risk factors and the category variable to classify whether dementia or not, and it does not carry out the longitude study. Future studies could fill up all 12 risk factors, and carry out a longitude study to

explore the dynamic influence on dementia after the cross-sectional study found the immediate risk factors.

References

[1] C. R. Beam, C. Kaneshiro, J. Y. Jang, C. A. Reynolds, N. L. Pedersen, and M. Gatz. Differences between women and men in incidence rates of dementia and alzheimer's disease. *Journal of Alzheimer's Disease*, 64(4):1077–1083, 2018.

[2] S. Bleeker, H. Moll, E. a. Steyerberg, A. Donders, G. Derksen-Lubsen, D. Grobbee, and K. Moons. External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology*, 56(9):826–832, 2003.

[3] E. M. Crimmins, J. K. Kim, K. M. Langa, and D. R. Weir. Assessment of cognition using surveys and neuropsychological assessment: the health and retirement study and the aging, demographics, and memory study. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(suppl_1):i162–i171, 2011.

[4] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with bayesian networks: A bootstrap approach. *arXiv preprint arXiv:1301.6695*, 2013.

[5] K. Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.

[6] M. I. Jordan. Graphical models. *Statistical science*, 19(1):140–155, 2004.

[7] K. B. Korb and A. E. Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

[8] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990.

[9] G. Livingston, A. Sommerlad, V. Orgeta, S. G. Costafreda, J. Huntley, D. Ames, C. Ballard, S. Banerjee, A. Burns, J. Cohen-Mansfield, et al. Dementia prevention, intervention, and care. *The Lancet*, 390(10113):2673–2734, 2017.

[10] K. G. Moons, A. P. Kengne, M. Woodward, P. Royston, Y. Vergouwe, D. G. Altman, and D. E. Grobbee. Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, 98(9):683–690, 2012.

[11] P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. *Encyclopedia of database systems*, 5:532–538, 2009.

[12] D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[13] M. Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(3), 2010.

[14] R. J. Stevens and K. K. Poppe. Validation of clinical prediction models: what does the "calibration slope" really measure? *Journal of clinical epidemiology*, 118:93–99, 2020.

[15] E. W. Steyerberg and Y. Vergouwe. Towards better clinical prediction models: seven steps for development and an abcd for validation. *European heart journal*, 35(29):1925–1931, 2014.

[16] D. H. Taylor Jr, T. Østbye, K. M. Langa, D. Weir, and B. L. Plassman. The accuracy of medicare claims as an epidemiological tool: the case of dementia revisited. *Journal of Alzheimer's Disease*, 17(4):807–815, 2009.

[17] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS conference*, volume 2, pages 376–380. St. Augustine, FL, 2003.

[18] R. Wittenberg, B. Hu, L. Barraza-Araiza, and A. Rehill. Projections of older people with dementia and costs of dementia care in the united kingdom, 2019–2040. *London: London School of Economics*, 2019.

[19] S. Yaramakala and D. Margaritis. Speculative markov blanket discovery for optimal feature selection. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4–pp. IEEE, 2005.

Appendices

A the code

The code of this report: code The latex of this report: latex

B Appendix for EDA

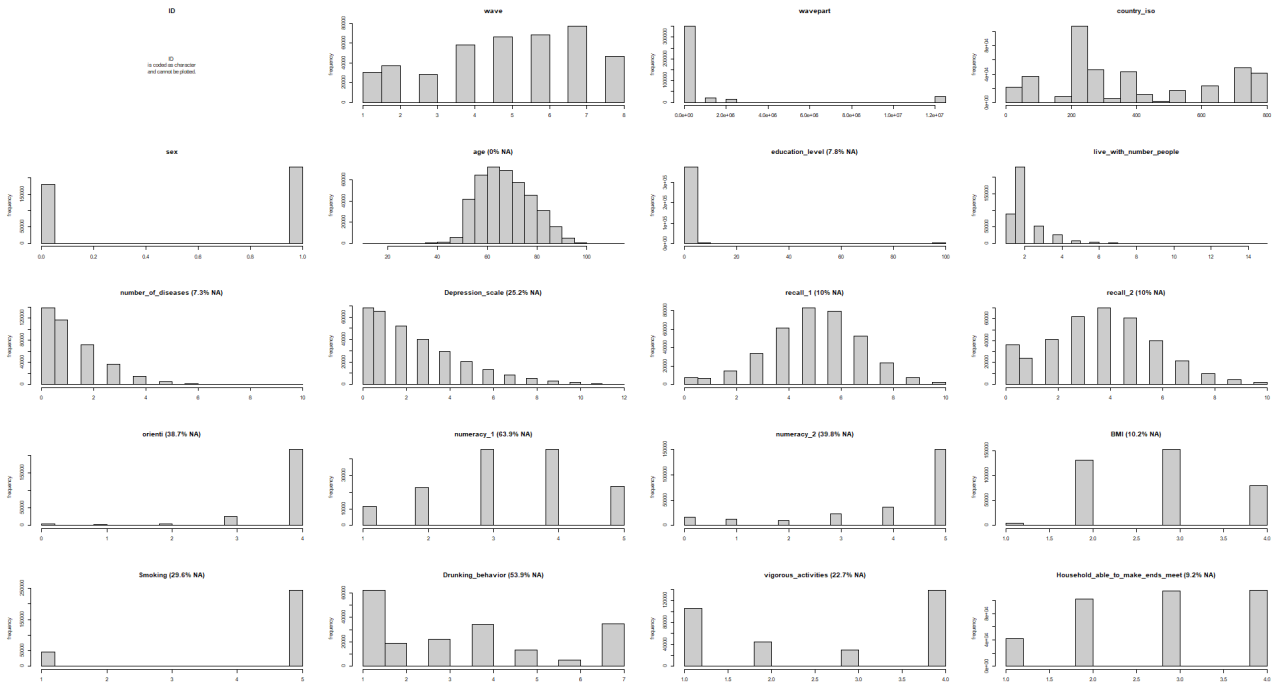

Figure 13: this is the plot of variables before EDA.

For sex, BMI, Smoking, and HAM variables, we need to change to factor variables.

For the Age variable, we only want to explore the people who are older than 60, people were grouped every five years until the people older than 90. The reason is that there are few people older than 90 from the figure 13.

For LWNP variable, we find there are less people live with more than 2 other people from the figure 13, so combine them as $(2, 13]$.

For NODs, we find there are less people have more than 2 diseases from the figure 13, so combine them as $[3, 11)$.

For DK, I decided to combine 1 and 2, 3 and 4, 5 and 6. The reason is that 1 and 2 means never drinking or less than once a month, 3 and 4 means one or twice a month or a week, 5, 6 and 7 are drinking more than 3 times a week. However, there are fewer people in 5 and 6, but many in 7 from the figure 13. Thus, we choose to combine 5 and 6 and leave 7 alone. Therefore, we divide the variable into 4 parts: $[1, 3)$, $[3, 5)$, $[5, 7)$, and $[7, 8)$.

For VA, we decided to change to a binary variable, $[1, 4)$ means to do some activities, $[4, 5)$ means hardly ever. The reason is that there are fewer people in 2 and 3 from the figure 13, and it does not necessary to separate this variable into multiple categories.

For EDUL, I decided to combine 0 and 1, 3 and 4, 5 and 6. Therefore, I divide the variable into 4 parts: $[0, 2)$, $[2, 3)$, $[3, 5)$ and $[5, 7)$.

For countries variable, this variable only appears in the multiple countries study. In this study, we choose 9 countries. We set 0 represent developing countries include (Bulgaria, Croatia, and Lithuania), 1 to represent developed countries include (France, Switzerland, and Germany) and 2 to represent developed countries in the Mediterranean include (Italy, Spain, and Greece).

The following plot is the histogram for each variable after EDA.

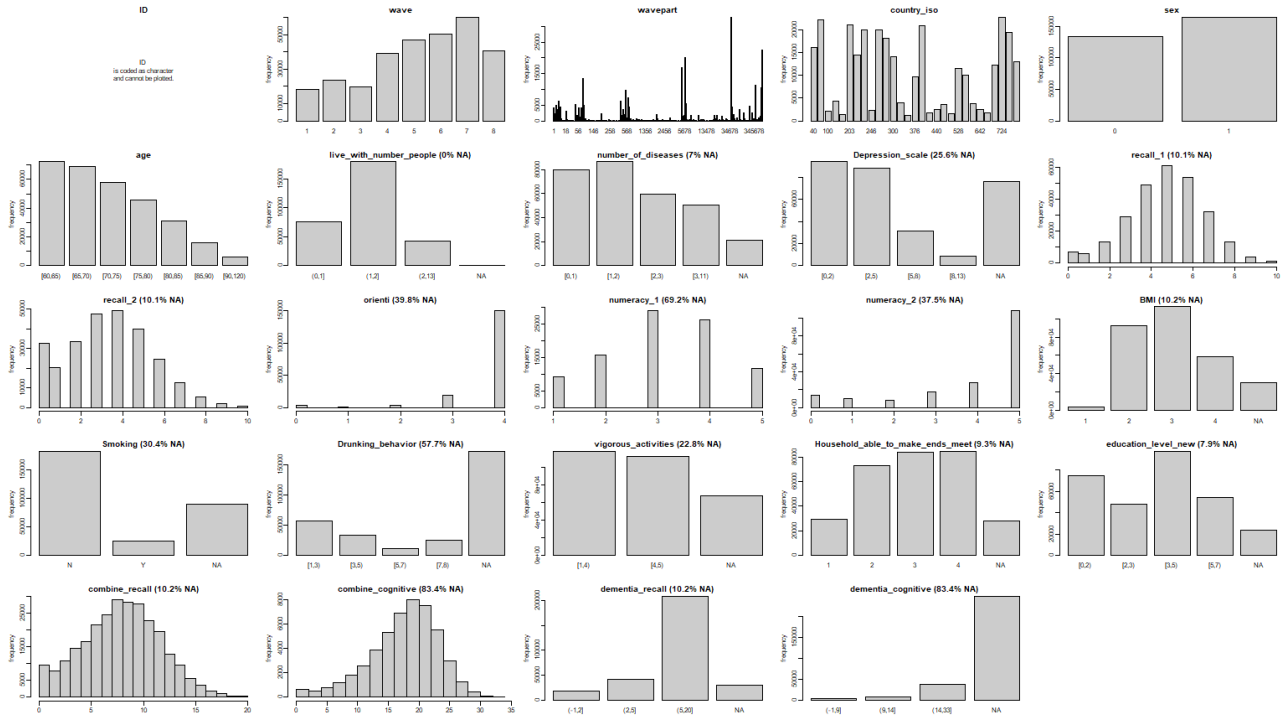

Figure 14: this is the plot of variables after EDA.

C Appendix for missing values

Missing pattern for different waves.

D Appendix for word count

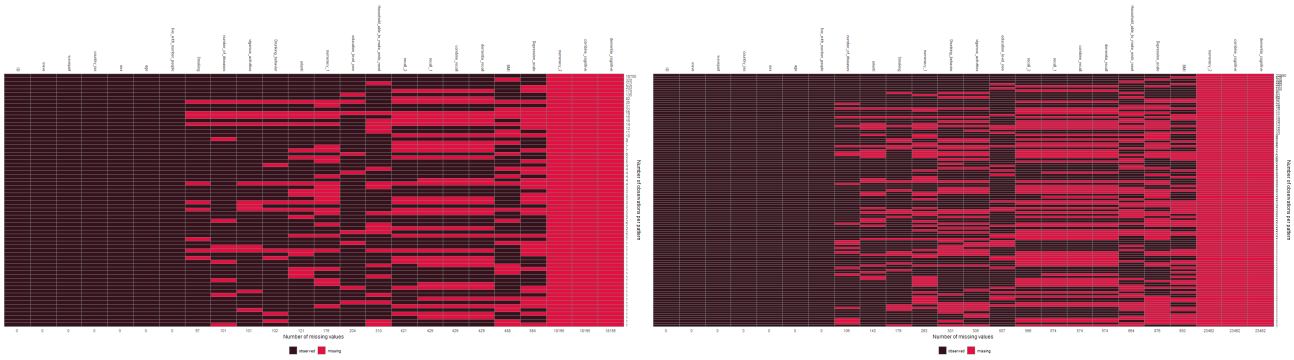

Figure 15: missing pattern for wave 1 (left) and 2 (right). Both plots are point that combine_cognitive is complete missing.

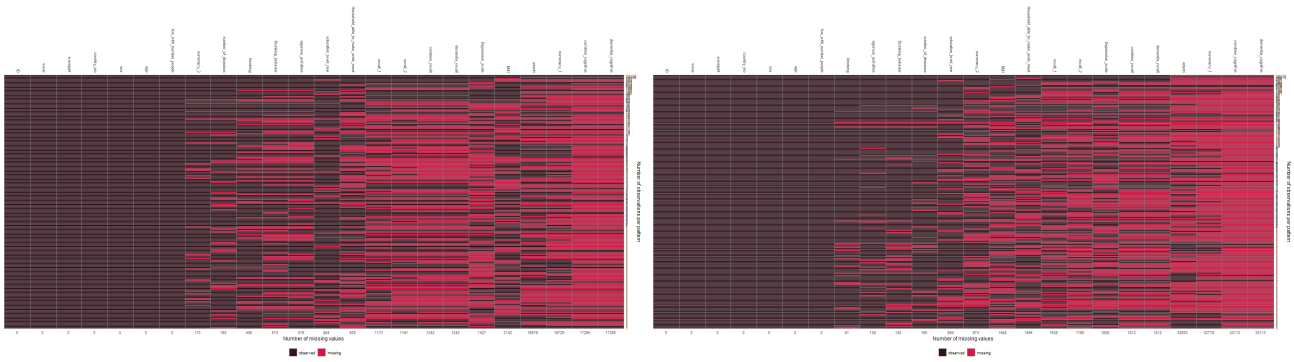

Figure 16: missing pattern for wave 4 (left) and 5 (right). Both plots are point that no variable is complete missing. The wave is a little better than wave 5.

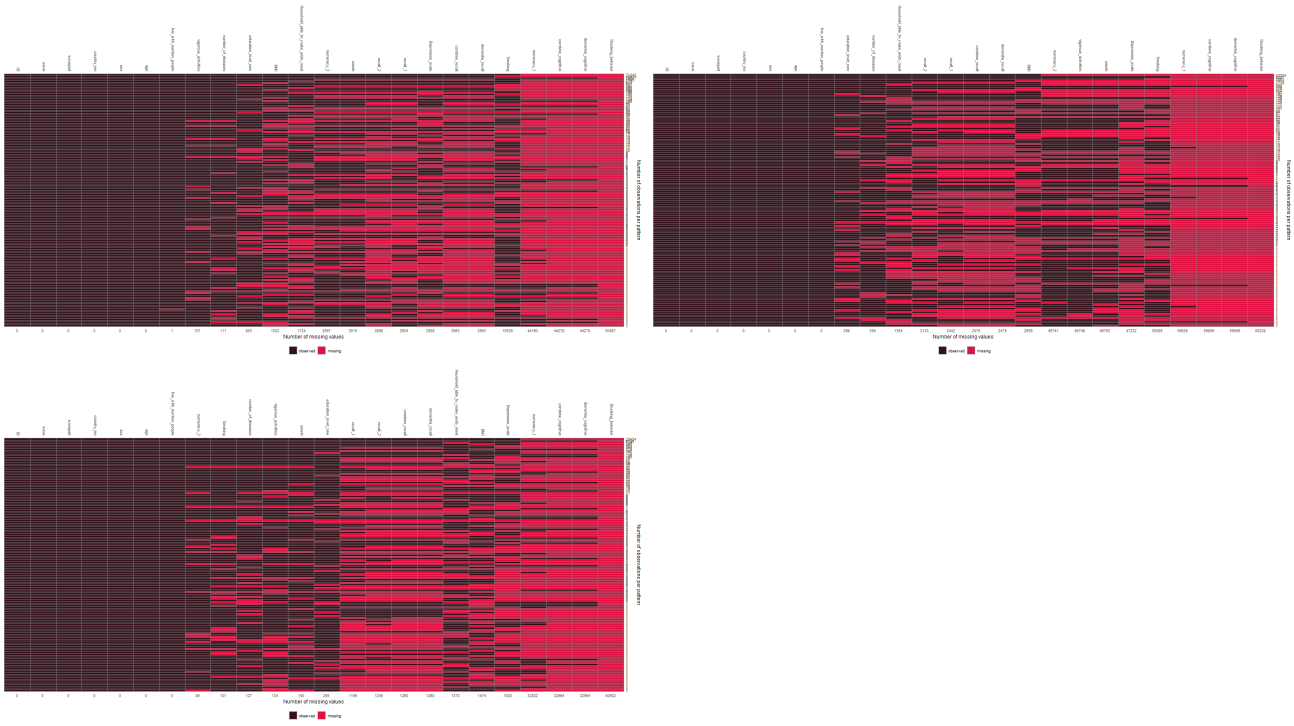

Figure 17: missing pattern for wave 6 (left), 7 (right) and 8 (bottom). All plots are point that drinking_behavior is complete missing.

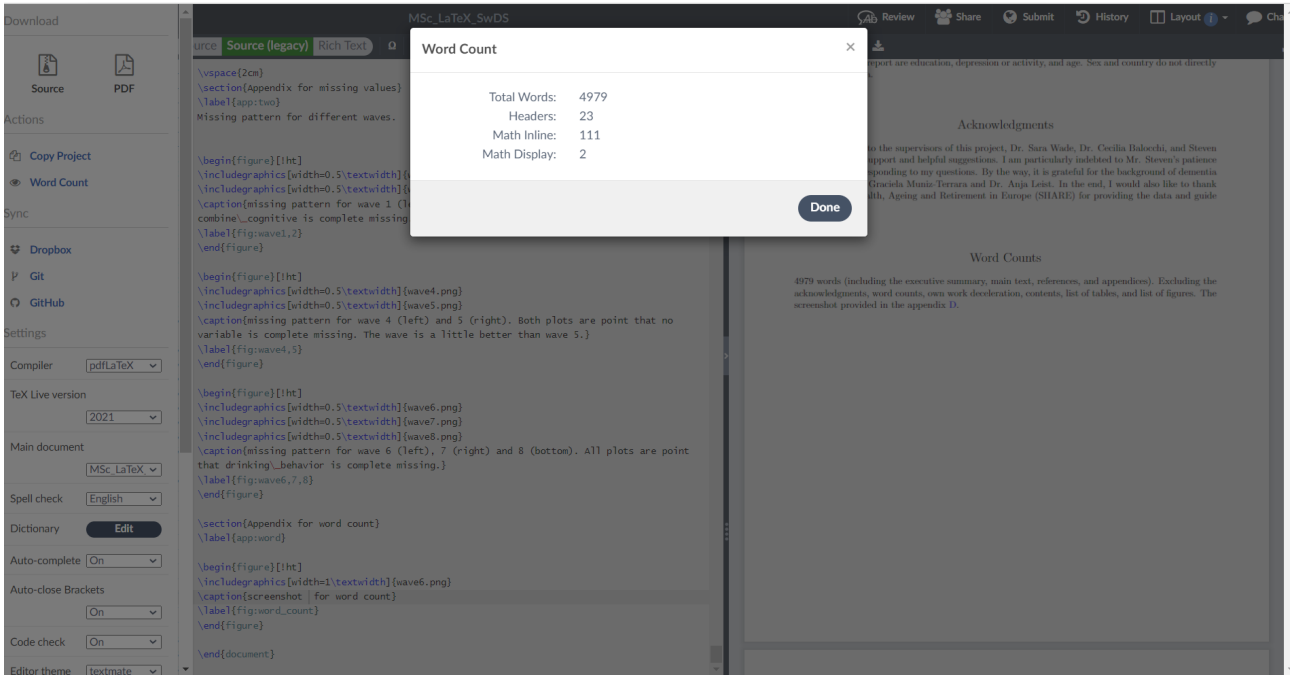

Figure 18: screenshot for word count