The School of Mathematics

Analysis on Dementia Risk Factors Based on Bayesian Networks

by

Yile Shi

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

June 2021

Supervised by
Dr. Sara Wade and Dr. Cecilia Balocchi

Executive Summary

Background: Dementia, a major international public health concern, seriously affects people's lives. It is reported that 50 million people around the world had been diagnosed with dementia in 2019, and this number is estimated to reach 152 million by 2050 [1]. Currently, the cure of dementia is still absent, hence effective prevention strategies become critical to reduce the risk of dementia and lessen its burden on society.

Research question: This research aims to shed light on the potential risk factors that affects individual's cognitive function, and further generate knowledge to inform the development of lifestyle interventions for dementia risk reduction. We consider 13 factors: age, gender, country, education, drinking behaviour, smoking, obesity, physical activity, chronic disease, working status, household finance, social connection and depression.

Data: We use a subset of easySHARE dataset from Survey of Health, Ageing and Retirement in Europe, which contains 57310 subjects recorded in 2013.

Methods: We construct Bayesian networks where the response of interest is individual's cognitive score. Modified network based on Hill-climbing algorithm is selected due to best performance through 10-fold cross-validation. Further, we use the network to estimate model parameters and plot line graphs to explore the trends of conditional probabilities of cognitive score in different groups.

Results: Age, gender, country, education and depression are shown to have significant effects on individual's cognitive impairment. Generally, age and depression have negative influence where individual's cognitive impairment gets more severe as age or depression level increases. Particularly, cognitive impairment seems to aggregate for individuals over 65 years or experiencing severe depression (figure 1). A positive effect of education level on cognition is observed, where people with higher education background usually have lower dementia risk. For the binary factor country, individuals living in countries with low average gross domestic product are more likely to suffer from dementia. Besides, female groups have higher dementia risk than males.

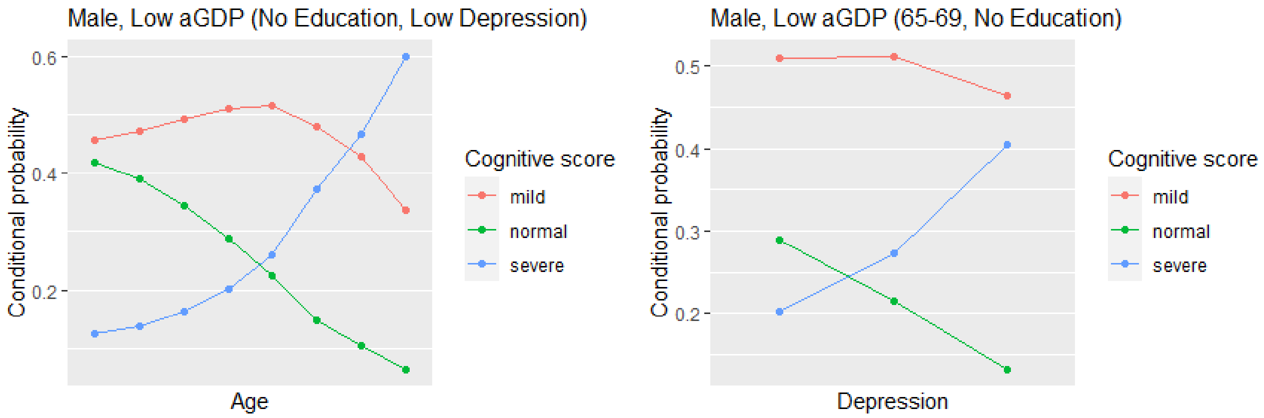


Figure 1: Examples of line graphs of cognitive impairment over age (left) and depression (right)

Acknowledgments

I am sincerely grateful to the supervisors of this project, Dr. Sara Wade and Dr. Cecilia Balocchi, as well as PhD supporter Steven Soutar, for their advice and guidance. I would also like to thank SHARE for providing the data and background information used in the research.

This report uses data from SHARE Waves 1 [2], 2 [3], 3 [4], 4 [5], 5 [6], 6 [7], 7 [8], and 8 [9]. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782, SHARE-COVID19: GA N°101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

This report also uses data from the generated easySHARE data set [10]. The easySHARE release 8.0.0 is based on SHARE Waves 1, 2, 3, 4, 5, 6, 7 and 8.

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: Yile Shi
Matriculation Number: s2168022
Title of work: Analysis on Dementia Risk Factors Based on Bayesian Networks

We confirm that all this work is my own except where indicated, and that We 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 We have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Complied with any other plagiarism criteria specified in the Course handbook

We 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:

Yile Shi

Date: 2022/6/30

Contents

1 Introduction

Dementia, a clinical state characterized by loss of function in multiple cognitive domains, becomes a serious public health concern worldwide [11]. Someone develops dementia every 3 seconds and current annual cost of dementia is estimated at 1 million dollars, which is set double in 2030 [1]. Because of the absence of effective treatment, prevention strategy to reduce dementia risk becomes an active research topic. This report aims to contribute to this topic by identifying potential risk factors behind dementia diagnosis. We expect that a better understanding of the influence of factors on cognitive impairment can help to inform the development of lifestyle interventions for dementia risk reduction.

We apply exploratory data analysis on easySHARE dataset from Survey of Health, Ageing and Retirement in Europe (SHARE) and extract a sample of 57310 individuals recorded in 2013. Based on the subset, we build Bayesian networks and select the one with the best performance. Our goal is to find, among the following factors, the most likely to be relevant to cognitive decline: age; gender; country; education; drinking behaviour; smoking; obesity; physical activity; chronic disease; working status; household finance; social isolation; and depression.

2 Exploratory Data Analysis

EasySHARE dataset is a combination of 8 waves of SHARE data recorded from 2004 to 2020, consisting of over 420k observations and 107 variables across various topics. It is inefficient to use all variables for modelling. On the other hand, from the summary of each column, we observe that the proportions of missing values in some columns are relatively high. For example, the variable `books_age10` is only available in wave 3 and 5, hence a large amount of missing values is shown in figure 2. Models using these features could loss much information on the data and result in inaccurate results. Thus, for model efficiency and accuracy, exploratory data analysis is necessary to explore the distribution of variables. We apply feature selection to drop redundant variables and feature transformation to define appropriate risk factors and response. Furthermore, as a result of cross-sectional research on a single wave, we extract a subset of data including these variables in a specific wave.

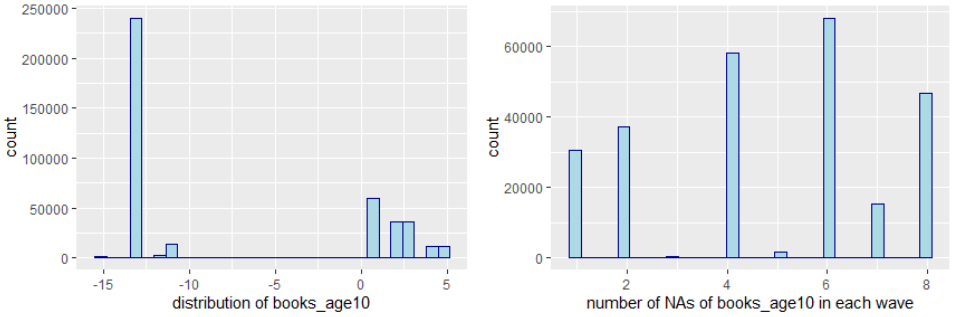


Figure 2: Information of `books_age10`

2.1 Feature Selection

2.1.1 Risk Factors

Referring to literatures of Crimmins et al (2011) [12], Beam et al (2018) [13] and Livingston et al (2020) [14], we determine modifiable risk factors in 13 domains, including **age**, **gender**, **country**, **education**, **drinking behaviour**, **smoking**, **obesity**, **physical activity**, **depression**, **social connection**, **chronic disease**, **working status** and **household finance**. For each factor, we follow the variable descriptions in easySHARE data guide [15] and select relevant variables following the rules below:

- **Variables for the same risk factor should have less information overlapping**.

  Some risk factors have multiple relevant features. An appropriate combination of these variables can represent this factor properly. Nevertheless, combination using redundant variables with great overlaps possibly lead to model inefficiency and bias. For example, `euro1` to `euro12` are binary variables indicating if respondents had 12 specific mental feelings in the past respectively, while `eurod` is a composite scale of them. Thus, `eurod` is enough to represent the depression level.

- **Variables should contain missing values as few as possible.**

  Although constrain-based algorithms can deal with missing values automatically when constructing Bayesian networks, algorithm evaluation through cross-validation requires non-missing dataset. Using inconsistent datasets between algorithms will cause great bias. On the other hand, most imputation methods based on other columns are not suitable in this research, which probably increases the model errors when variables are not strongly correlated. Hot deck imputation using values from the same column in other waves also causes great bias as values in most variables probably change over the time. In this case, we have to drop incomplete observations from the dataset, so variables with fewer missing values are preferred as they keep the maximum information of the original data.

  For example, considering one's social isolation, `mar_stat` and `partnerinhh` are both fine to represent this factor, while only one of them can be selected due to the former rule. Comparing the histograms 3, we choose `partnerinhh` without missing values.

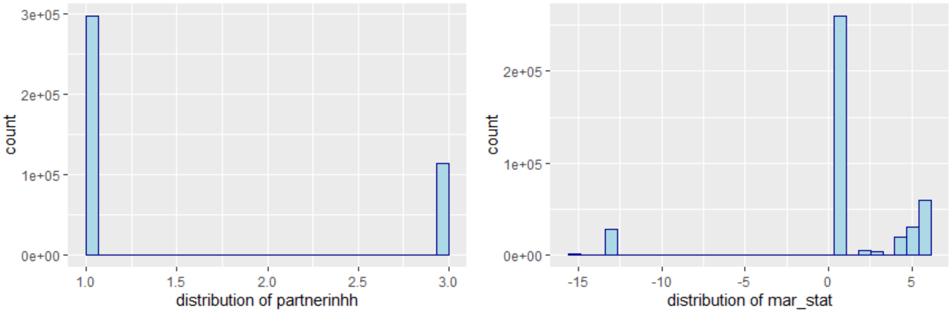


Figure 3: Distributions of `partnerinhh` and `mar_stat`

Table 1 displays 13 variables selected to represent the risk factors based on the rules above.

Risk Factor	Variables	Description
Age	`age`	age at interview
Gender	`female`	gender of respondents
Country	`country`	country identifier
Education	`isced1997_r`	ISCED-97 coding of education status
Drinking behaviour	`br010_mod`	frequency of alcohol drinking
Smoking	`smoking`	if respondents smoke at present
Obesity	`bmi2`	categorized body mass index
Physical activity	`br015_`	the frequency of vigorous activities
Depression	`eurod`	scale of current depression
Social connection	`partnerinhh`	if respondents live with spouse/partner
Chronic disease	`chronic_mod`	number of chronic diseases
Working status	`ep005_`	current job situation
Household finance	`co007_`	confidence on household income

Table 1: Risk factors and corresponding variables

2.1.2 Cognitive Score

Next, we explore the variables corresponding to the model output. EasySAHRE dataset does not record diagnosis of dementia (e.g. Alzheimer's disease) in all waves. Instead, it contains the following indices to describe respondents' cognitive function:

- `recall_1`: the number of words recalled in the first trial of the word recall task, ranging from 0 to 10.
- `recall_2`: the number of words recalled in the delayed word recall task, ranging from 0 to 10.
- `orienti`: orientation of date, month, year and day of week, ranging from 0 (good) to 4 (bad).
- `numeracy_1`: information on the Mathematical performance, ranging from 1 (bad) to 5 (good).
- `numeracy_2`: information on the second test on Mathematical performance, ranging from 0 to 5.

Following Crimmins' [12], we consider creating a composite cognitive score as the response variable in subsequent modelling, using these indices as a proxy for dementia severity. We visualize the distributions of these features in figure 4:

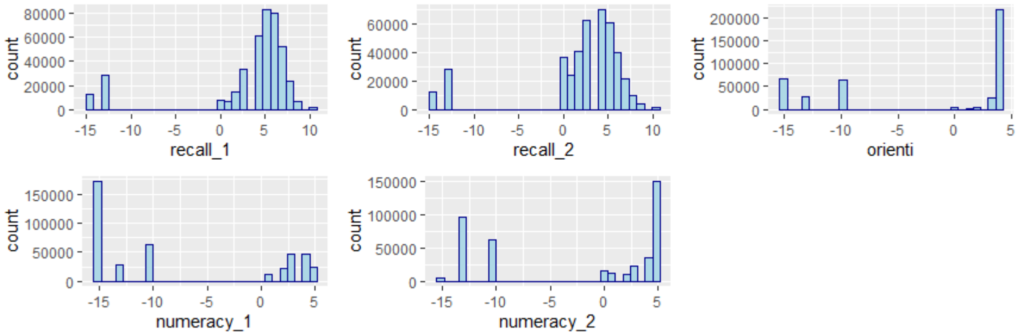


Figure 4: Distributions of cognitive function indices

A high amount of missing values is observed in `orienti`, `numeracy_1` and `numeracy_2`. According to easySHARE guide, `orienti` and `numeracy_1` have a large number of missing values in wave 4 to 8, while `numeracy_2` is only available for respondents of wave 4 to 8 who already participated in one panel waves. Thus, for numeracy scores, only one of both is recorded in most observations. As mentioned, we do not apply imputation methods here and columns with fewer incomplete data are preferred.

After variable selection, we consider extracting data of a specific wave from easySHARE dataset. Wave with more observations and higher proportion of complete data in each column is ideal. We compare data of aforementioned variables in 8 waves. For waves having over 20% rows missing in `orienti`, we drop this column when comparing. For `numeracy_1` and `numeracy_2`, we pick either of them with more complete data and drop the other one. As a result, wave 5 is selected.

Furthermore, we drop the observations with `age` less than 50 as they are not the people of interest. We also notice that their exists a relatively small proportion (<0.2%) of observations with values over 90 in `isced1997_r` or `ep005_`. According to easySHARE guide, these values represents the option *others* and we also drop these observations. As a result, 57310 rows are kept for modelling.

2.2 Feature Transformation

Up to now, we have already extracted a subset including 17 variables in wave 5 from easySHARE dataset. Before subsequent modelling, we apply the following feature transformation steps on this subset:

1. **Rename the variables for risk factors.**

   We rename most variables with the names of their corresponding risk factors to represent them, instead of abbreviations or numerical coding.

2. **Create cognitive score as model response.**

   We create a composite cognitive score as the model response using corresponding indices. Since `orienti` and `numeracy_1` has already been dropped, the cognitive score of a respondent is defined as the sum of `recall_1`, `recall_2` and `numeracy_2`, ranging from 0 to 25. The higher the score is , the better cognitive function the respondent has.

3. **Discretize continuous nodes.**

   Due to the property of Bayesian Networks that discrete nodes only have discrete parents, we discretize continuous nodes so that they can be properly used. Referring to Livingston's (2020), the cognitive score is cut into 3 bins, based on its 10% and 40% quantiles, to represent severe impairment, mild impairment, and cognitively normal. Continuous nodes including age, chronic diseases and depression are discretized to ensure that each categories contains sufficient data in case of model overfitting. Figure 5 shows the distributions of cognitive score and age, before and after discretization.

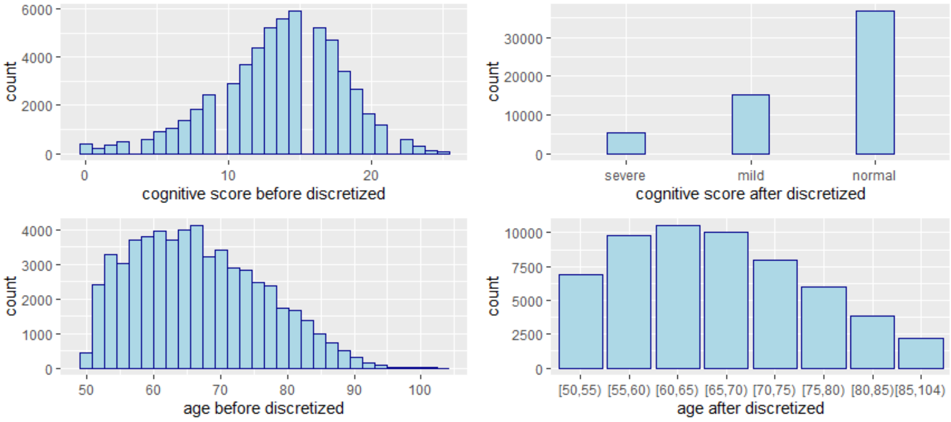


Figure 5: Distributions of cognitive score and age

   Besides, country is also discretized to a binary variable. We refer to the average gross domestic product (aGDP) of relevant countries in 2013 when wave 5 was recorded and set a threshold of 40k dollars. Table 2 shows countries with identifiers in each group. Figure 6 displays the distribution of country before and after discretization.

1- High aGDP (≥\$40k)	0 - Low aGDP (<\$40k)
Austria (11)	Spain (15)
Germany (12)	Italy (16)
Sweden (13)	Israel (25)
Netherlands (14)	Czech (28)
France (17)	Slovenia (34)
Denmark (18)	Estonia (35)
Switzerland (20)	
Belgium (23)	
Luxembourg (31)	

Table 2: Countries categorized based on aGDP in 2013

4. **Recategorize some discrete factors.**

   Some categories of discrete factors, including education, drinking behaviour and physical activity, do not contain enough data, which leads to model overfitting. As a remedy, we recategorize these variables by combining some levels to ensure the number of observations in each stratification is sufficient. Figure 7 shows the distributions of drinking behaviour, before and after recategorized.

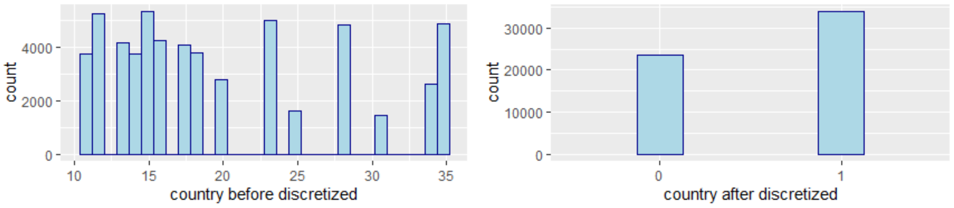


Figure 6: Distribution of country

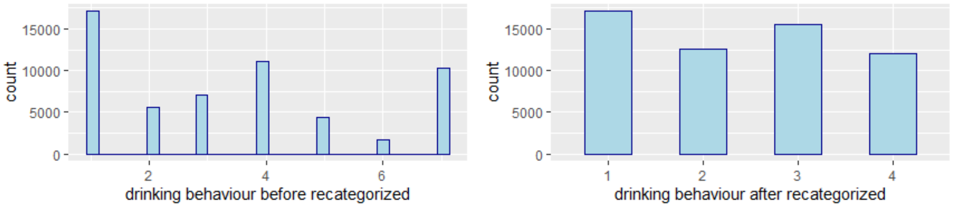


Figure 7: Distributions of drinking behaviour

After feature transformation, the subset is ready for subsequent modelling. Table 3 shows the levels of risk factors and cognitive score.

Risk factors	Levels
Age	1: 50-54; 2: 55-59; 3: 60-64; 4: 65-69; 5: 70-74; 6: 75-79; 7: 80-84; 8: $\geq$85
Gender	0: male; 1: female
Country	0: low aGDP; 1: high aGDP
Education	0: no education; 1: primary education; 2: lower secondary education; 3: secondary education; 4: post-secondary and tertiary education
Drinking behaviour	1: never; 2: sometimes; 3: often; 4: almost everyday
Smoking	0: no, 1: yes
Obesity	1: underweight; 2: normal; 3: overweight; 4:obese
Physical activity	1: more than once a week; 2: one to four times a month; 3: hardly ever, or never
Depression	1: low; 2: medium; 3: high
Chronic diseases	1: no disease; 2: one or two diseases; 3: over two diseases
Working status	1: retired; 2: employed; 3: unemployed; 4: permanently sick or disabled; 5: homemaker
Household finance	1: with great difficulty; 2: with some difficulty; 3: fairly easily; 4: easily
Cognitive score	1: severe; 2: mild; 3:normal

Table 3: Levels of risk factors and cognitive score

3 Implementation

In this section, we introduce the Bayesian network and algorithms used for modelling. Moreover, we explain the evaluation of networks, and further the modification on the selected network.

3.1 Bayesian Network

The exposition in this subsection follows that in Scutari (2010) [16].

A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). The graph is denoted by $\mathcal{G} = (V, E)$, where V is the node set consisting of variables of interest and E is the edge set. Each edge in E represents a direct dependence between two noes. For example, edge "A → B" means that node B depends on A. In this case, A is called the parent, while B is called the child. As name suggests, the graph supposed to be acyclic. Cycles such as "A → B → A" are not permitted. Table 4 shows the 3 types of Bayesian networks:

Type	Feature	Common choices of local distributions
Discrete	Only has discrete variables	Multinomial, Binomial, Poisson
Continuous	Only has continuous variables	Multivariate normal, Student-t, Beta
Mixed	Has both discrete and continuous variables	-

Table 4: Types of Bayesian networks

Bayesian network analysis starts with structure learning, which has two categories of algorithms:

- **Constraint-based algorithms**: these algorithms learn the network structure by analysing the probabilistic relations entailed by the Markov property of Bayesian networks with conditional independence tests (e.g. mutual information) and then constructing a graphs satisfying the corresponding d-separation statements. The common choices could be *Peter-Clark (PC)* algorithm, and *Incremental Association (IAMB)* algorithm.

- **Score-based algorithms**: these algorithms assign a score (e.g. Bayesian Information Criteria) to each candidate Bayesian network and try to maximize it with some heuristic search algorithm. *Hill-climbing (HC)*, based on greedy search algorithms, is the common choice.

Next, a blacklist and a whitelist are defined, based on general information and expert knowledge, to avoid or ensure some edges in the network. We define the blacklist following the rules below:

- No arrow starts from the response node, cognitive score.

- No arrow points to nodes of respondents' natures, i.e. age, gender or country.

- As an early-life factor, education is only possibly influenced by age, gender and country.

We define the whitelist with two edges "working status → household finance" and "social isolation → depression" based on our common sense.

3.2 Evaluation of Algorithms

We now evaluate the performance of structure learning algorithms including *PC*, *IAMB* and *HC*. For each algorithm, we calculate the average loss for the response node, cognitive score, through 10-fold cross-validations on the extracted subset. Classification error is implemented as the loss function, because most loss functions, including mean squared error, are not available in discrete networks.Values of cognitive score are predicted using the information present in its local distribution from its parent nodes. Lower value of the loss indicates better performance. Considering the random effect of cross-validation, we repeat computing average loss of each algorithm with different random seeds. As shown in table 5, network based on HC over-performs the other two algorithms. Therefore, HC is selected for further modification and estimation. Figure 8 displays the initial Bayesian Network based on HC.

algorithm	seed = 1	seed = 2	seed = 3	seed = 4	seed = 5
PC	0.3285017	0.3273365	0.3285484	0.3274023	0.3276248
IAMB	0.3231999	0.3242536	0.3238454	0.3234789	0.3229207
HC	0.3226488	0.3224045	0.3230326	0.3227360	0.3227883

Table 5: Average loss of algorithms after 10-fold cross-validation with different random seeds

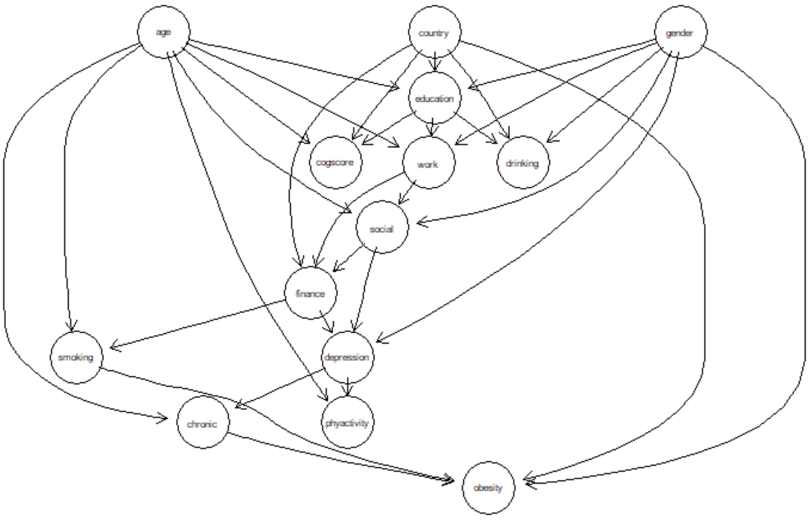


Figure 8: Initial Bayesian network created using HC algorithm

3.3 Model Modification

The network above may miss some edges of interest as algorithms do not exhaust all possible edges. Besides, casual effects indicated in some edges violates our general knowledge, although they are statistically significant. Thus, the network need further modification to make it more reasonable. Unlike constraint-based algorithms, HC doesn't require conditional independence test before any modification. We compare model performance after all modifications are finished.

First, we drop the edge "household finance → smoking" as we do not believe the financial status of the household affects one's smoking habit directly. Similarly, edges "education → drinking behaviour" and "depression → physical activity" are dropped. Besides, we reverse the edge "depression → chronic disease" following the logic that physical diseases influence individual's mental health.

Next, we add the edge "gender → cognitive score" into the network as gender is believed to have significant effect on respondents' cognition according to [13]. Moreover, since all risk factors are assumed to have direct or indirect effect on individual's cognitive function through at least one path, we add the edges "depression → cognitive score", "obesity → depression", "physical activity → obesity" and "drinking behaviour → obesity" to ensure the assumption holds.

We calculate the average loss of modified networks through 10-fold cross-validations, and compare the results with the initial network. We repeat the computation with different random seeds to reduce the random effect of cross-validations. The results are shown in table 6:

seed	before	after
1	0.3226488	0.3191122
2	0.3224045	0.3198102
3	0.3230326	0.3190773
4	0.3227360	0.3192692
5	0.3227883	0.3184842

Table 6: Average classification error of the network before and after modification

As a result, the modified network performs better and is selected for subsequent estimation and interpretation.

4 Results

The modified Bayesian network based on Hill-climbing algorithm is determined as the final network, shown in figure 9. As the parents of cognitive score, **age**, **gender**, **country**, **education** and **depression** tend to have significant effects on respondents' cognitive function.

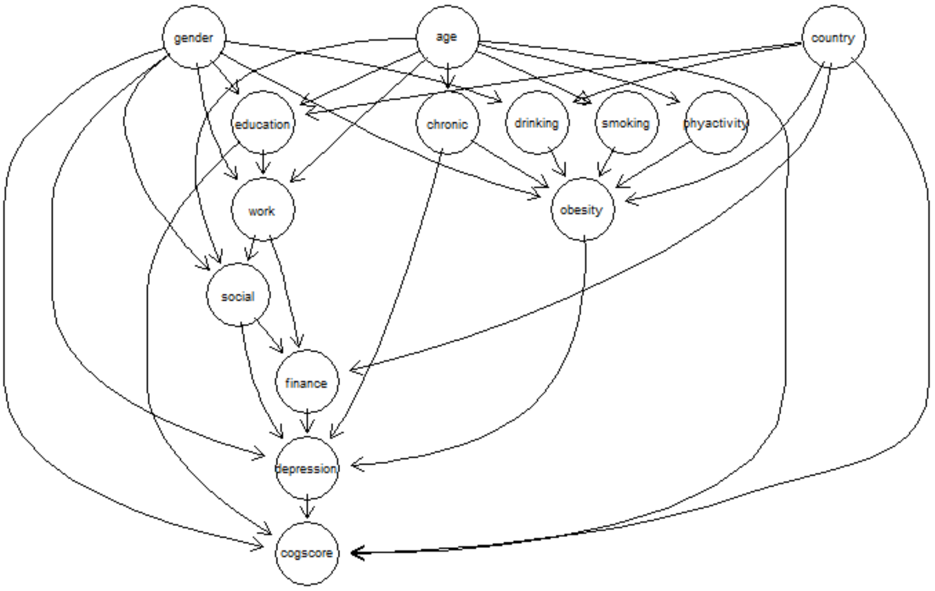


Figure 9: Final Bayesian network

Next, we estimate the parameters of local distributions of these parent nodes. We use the classical Bayesian posterior estimator with imaginary sample size of 10000, which provides smoother and more robust estimates, comparing with the alternative method using maximum likelihood estimation. We plot line graphs to explore the trends of the conditional probabilities of cognitive score, conditioning on its parent nodes. Since model overfitting happens in stratifications with sparse counts when conditioning on too many nodes, plots of the basic model are wiggly. As a remedy, we apply ordered logistic regressions of cognitive score, grouped by gender, over other parent nodes. Figure 10 displays the improvement of line graphs after using ordered logistic regression.

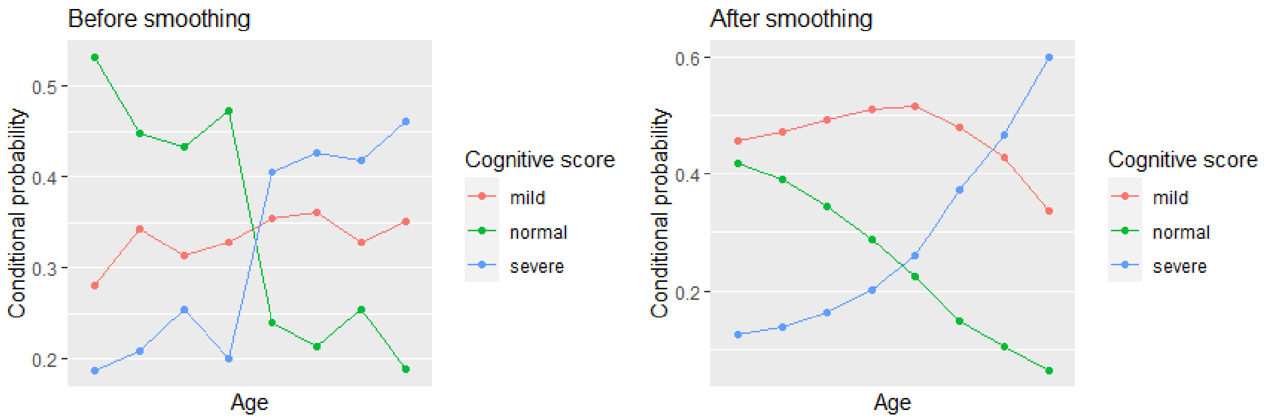


Figure 10: Line graphs for same data before and after smoothing

With smoother line graphs, we discuss the influence of parent nodes on dementia risk:

4.1 Age

Age shows a negative influence on individual's cognitive function from the line graphs. Generally, as respondents age, the conditional probability of having severe cognitive impairment increases, while the probability of being cognitively normal drops. This is consistent with our common sense. As people get older, they suffer from worse health condition and more loneliness, which could lead to cognitive decline. The path ”age → chronic diseases → depression → cognitive score” in the network supports this view.

We condition on people’s education and depression level. Given a specific gender group, we observe better cognitive function among respondents from countries with high aGDP in all age groups, as they have higher probability to be cognitively normal and lower probability to have severe cognitive impairment than those from countries with low aGDP. On the other hand, looking at individuals from the same country group, female respondents show higher chance to suffer from severe cognitive impairment than males, especially in higher age groups (e.g. over 85). Figure 11 displays the trends of cognitive score over age in different gender and country groups, among respondents with no education and low depression level.

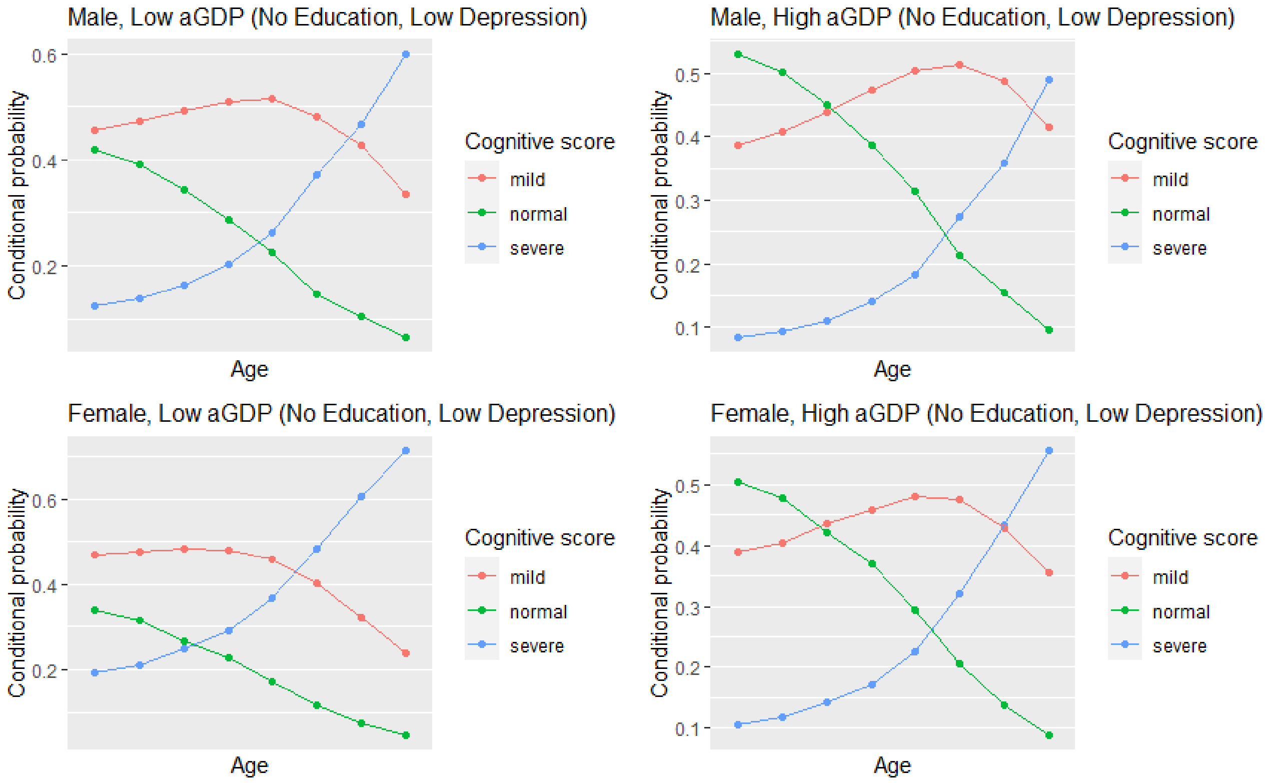


Figure 11: Trends of cognitive score over age (no education and low depression level)

Interestingly, we notice that most lines become steeper after 65 yeas old, indicating faster decline on people’s cognition. Thus, we think that 65 could be an important time point where individual’s cognitive decline aggregates. This is consistent with results from the study of Lee et al (2018) [17] in China. Furthermore, we particularly focus on the influence of other factors in the age groups around 65.

4.2 Education

Education level tends to have a positive effect on dementia risk reduction. According to line graphs, conditional probability of severe cognitive impairment decreases when respondents have higher education levels, and these people have higher chance to be cognitively normal. It has been proved that education, as a main contributor, stimulates people’s cognitive function in early life according to the work of Black et al (2018) [18]. Besides, individual’s education level might influence the work and further the income in midlife, which are also believed to have effects on cognition in other researches [14].

Similarly, we condition on age and depression level here. In each gender group, though the trends of lines are very similar, people from countries with low aGDP seems to have higher dementia risk, due to lower conditional probability of being cognitively normal and higher probability of having severe impairment on cognition. In each country group, again, female respondents could be more likely to suffer from cognitive impairment, particularly in low education levels. Figure 12 shows the trends of cognitive score over education levels in different gender and country groups, holding 65-69 yeas old and low depression level.

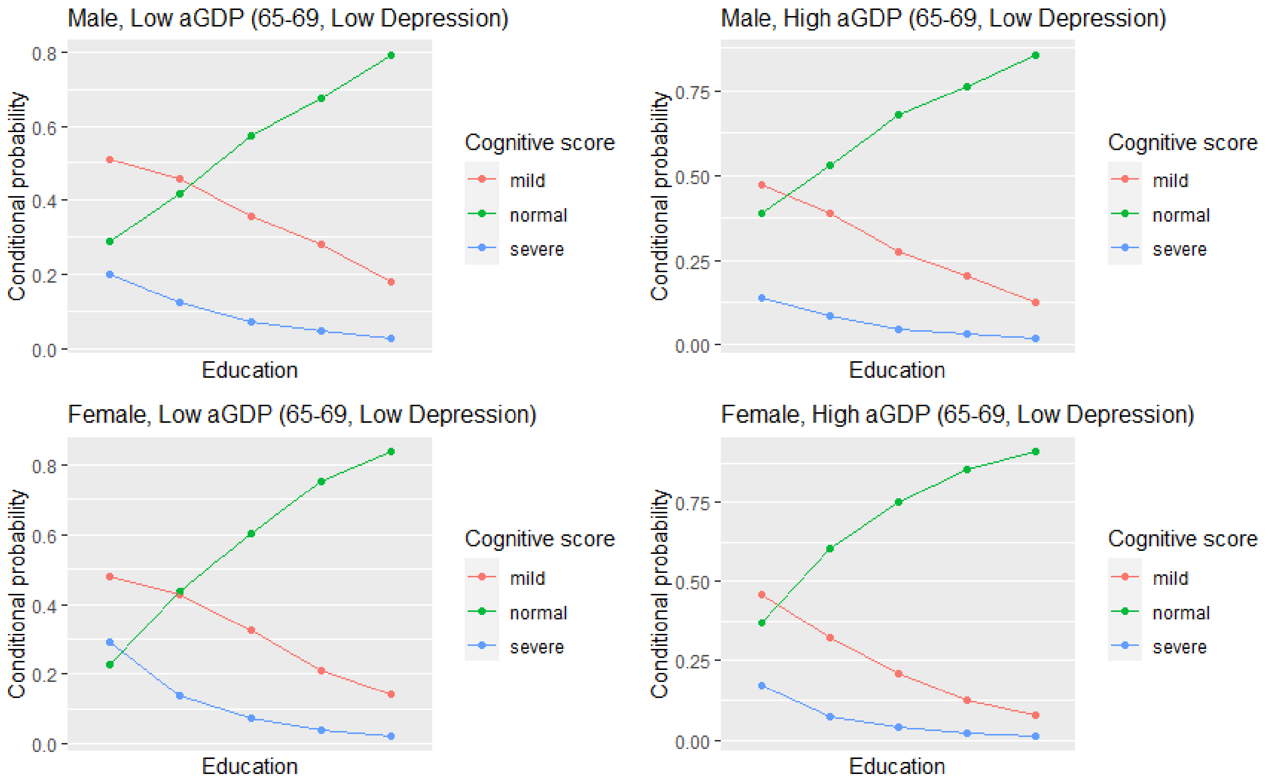


Figure 12: Trends of cognitive score over education (65-69 years old, low depression level)

4.3 Depression

Depression is found to aggravate individual's cognitive impairment. From the line graphs, we observe that as depression level increases, the conditional probability of normal cognition drops and people are more likely to suffer from severe cognitive impairment. As a part of the prodrome and early stages of dementia, depression is associated with various possible psychological or physiological mechanisms. This can be supported by edges corresponding to depression across many domains in the network including "social → depression", "household finance → depression, "chronic diseases → depression" and "obesity → depression". Poor health condition, low income and social isolation all tends to cause the feeling of depression, and further influence individual's cognition.

Given age, gender and education level, the conditional probability of normal cognition of respondents from countries with high aGDP is significantly higher than those from countries with low aGDP at each depression level. Accordingly, the probability of suffering from severe cognitive impairment in countries with high aGDP is much lower than it in the other stratification. Conditioning on the country group, we find that women possibly have higher dementia risk than men from the plots. Figure 13 shows the trends of the conditional probabilities of different cognitive function levels over depression, holding 65-69 age group and no education level.

Moreover, we notice that segments from medium to high depression level become much steeper than those from low to medium level. According to the ranges of depression levels, respondents with at least 5 negative feelings are determined with high depression. Thus, we guess it could be a threshold that people experiencing over 4 negative feelings have much higher dementia risk than others.

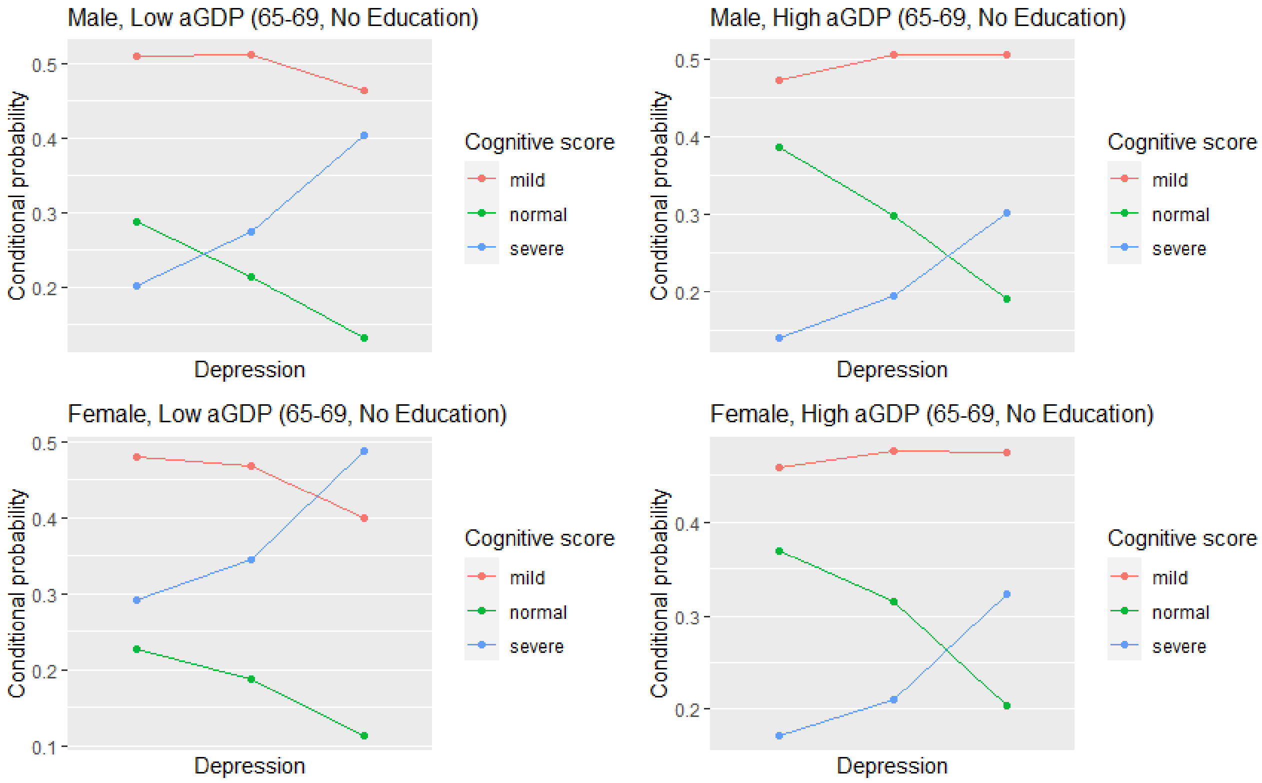


Figure 13: Trends of cognitive score over depression (65-68 years old, no education)

4.4 Gender

Gender shows consistent influence on cognitive function across analysis on age, education and depression that generally females have higher dementia risk than males, holding other factors constant. In particular, this difference is observed to be more significant in elder age groups. These results are consistent with the conclusions in [13]. A possible explanation for females having higher probabilities of cognitive impairment may be that women generally survive to longer ages than men, making the proportion of women become larger in elder groups. On the other hand, due to their longevity, female respondents are more likely to live alone. This could result in more severe social isolation, increases their depression level and finally affects their cognition, which is supported by the path ”female → social isolation → depression”. Besides, we believe that gender could affect individual’s cognitive function in other ways including the education level and obesity, from the network.

4.5 Country

Recall that we divide countries into two groups, based on their average gross domestic product in 2013. According to previous analysis, we conclude that respondents from countries with high aGDP tend to have lower dementia risk than those from countries with low aGDP. The influence of country can be various. From the network, we observe that country are significantly associated with nodes including education, obesity, drinking behaviour and household finance. Generally, countries with high aGDP probably have better welfare, which means they could provide higher-standard education and more advanced health care for their citizens than other countries, positively affecting the cognitive function. Moreover, provided with better welfare, people in these countries may live with more joy, hence they have lower depression level, which also reduces the probability of cognitive decline.

5 Conclusions

We construct a modified Bayesian network on a subset of easySHARE dataset to identify risk factors on dementia. As a result, age, gender, country, education and depression tend to have significant influence on individual's cognitive function. Specifically, the probability of severe cognitive impairment rises as age or depression level increases. In particular, individuals over 65 years old or experiencing high-level depression have much higher dementia risk. Education shows opposite influence on cognitive function that people with higher education levels have lower risk. Gender and country affect individual's cognition in a variety of ways. Generally, females have higher dementia risk than males; people from countries with low aGDP have higher probability to suffer from dementia.

There are some limitations of this research. First, although hearing loss, diabetes, and hypertension are believed to have important effect on cognitive impairment in [14], their influence could not be identified due to the lack of relevant columns in easySHARE data. We could only define a general factor "chronic diseases" to count the number of chronic diseases the respondents had and explore its possible influence on dementia risk instead. To determine the specific effects of these diseases, we need to access data with relevant variables. Second, our research belongs to cross-sectional analysis, which focuses on data at a specific time point among different individuals and obtains conclusions for the general population. In this case, our research is lack of the exploration on longitudinal effects of some risk factors on dementia risk among particular respondents over the time.

References

[1] Alzheimer's Disease International. World alzheimer report 2019: attitudes to dementia. *Alzheimer's Disease International: London*, 2019.

[2] A. Börsch-Supan. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 1*. Release version: 8.0.0. SHARE-ERIC. Data set, 2022. doi: 10.6103/SHARE.w1.800.

[3] A. Börsch-Supan. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2*. Release version: 8.0.0. SHARE-ERIC. Data set, 2022. doi: 10.6103/SHARE.w2.800.

[4] A. Börsch-Supan. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 3*. Release version: 8.0.0. SHARE-ERIC. Data set, 2022. doi: 10.6103/SHARE.w3.800.

[5] A. Börsch-Supan. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4*. Release version: 8.0.0. SHARE-ERIC. Data set, 2022. doi: 10.6103/SHARE.w4.800.

[6] A. Börsch-Supan. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5*. Release version: 8.0.0. SHARE-ERIC. Data set, 2022. doi: 10.6103/SHARE.w5.800.

[7] A. Börsch-Supan. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 6*. Release version: 8.0.0. SHARE-ERIC. Data set, 2022. doi: 10.6103/SHARE.w6.800.

[8] A. Börsch-Supan. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7*. Release version: 8.0.0. SHARE-ERIC. Data set, 2022. doi: 10.6103/SHARE.w7.800.

[9] A. Börsch-Supan. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 8*. Release version: 8.0.0. SHARE-ERIC. Data set, 2022. doi: 10.6103/SHARE.w8.800.

[10] A. Börsch-Supan and S. Gruber. *easySHARE*. Release version: 8.0.0. SHARE-ERIC. Data set, 2022. doi: 10.6103/SHARE.easy.800.

[11] Flávio Luiz Seixas, Bianca Zadrozny, Jerson Laks, Aura Conci, and Débora Christina Muchaluat Saade. A bayesian network decision model for supporting the diagnosis of dementia, alzheimer's disease and mild cognitive impairment. *Computers in biology and medicine*, 51:140–158, 2014.

[12] Eileen M Crimmins, Jung Ki Kim, Kenneth M Langa, and David 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.

[13] Christopher R Beam, Cody Kaneshiro, Jung Yun Jang, Chandra A Reynolds, Nancy L Pedersen, and Margaret 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.

[14] Gill Livingston, Andrew Sommerlad, Vasiliki Orgeta, Sergi G Costafreda, Jonathan Huntley, David Ames, Clive Ballard, Sube Banerjee, Alistair Burns, Jiska Cohen-Mansfield, et al. Dementia prevention, intervention, and care. *The Lancet*, 390(10113):2673–2734, 2017.

[15] Stefan Gruber, Christian Hunkler, and Stephanie Stuck. *Generating easySHARE: guidelines, structure, content and programming*. SHARE Working Paper Series (17-2014). Munich: MEA, Max Planck Institute for Social Law and Social Policy, 2014.

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

[17] Allen TC Lee, Marcus Richards, Wai C Chan, Helen FK Chiu, Ruby SY Lee, and Linda CW Lam. Association of daily intellectual activities with lower risk of incident dementia among older chinese adults. *JAMA psychiatry*, 75(7):697–703, 2018.

[18] Deborah Blacker and Jennifer Weuve. Brain exercise and brain outcomes: does cognitive activity really work to maintain your brain? *JAMA psychiatry*, 75(7):703–704, 2018.

Appendix

Programming part in this research is completed with R.

R packages:

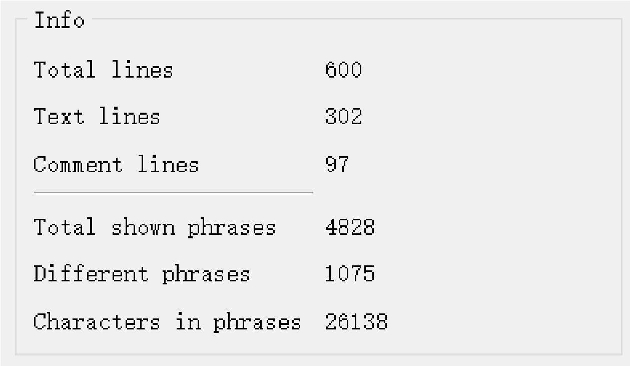
For EDA section, we use the `tidyverse` and `gridExtra` packages. We construct and estimate the networks with `bnlearn` package and display the network with `Rgraphviz` package. `MASS` package is used to fit the ordered logistic regression on cognitive score. All histograms and line graphs are generated with `ggplot2`.

R code:

The complete code is available via this Github repository.

Word Count

This report contains 4828 words, including executive summary, main text, references and appendix. The screenshot using `Analyse Text` function in TeXstudio is provided.



```
Info
Total lines             600
Text lines              302
Comment lines           97
Total shown phrases     4828
Different phrases       1075
Characters in phrases   26138
```

Figure 14: Word count in TeXstudio