Causal Discovery for Risk Factors and Diseases in Ageing Process

Yujia Gu

May 7, 2022

1 Introduction

1.1 Background

Population aging has been a worldwide topic in recent years. As reported by World Health Organization(WHO) in Oct. 2021, by 2030, 1 in 6 people in the world will be aged 60 years or over. At the same time, every person should have the opportunity to live a long and healthy life. Therefore, healthy aging becomes a vital problem in our society. As one of the opposites to health, diseases, especially non-communicable diseases (NCDs) such as heart disease, cancer and diabetes, contribute a lot to the loss of health and life. Many risk factors have been found that have a relationship with non-communicable disease, such as unhealthy diets, physical inactivity, exposure to tobacco smoke or the harmful use of alcohol[Ezzati and Riboli, 2013]. A question of great interest is the causal relationship between risk factors and the incidence of diseases during the aging process. Especially, we aim to figure out the change of this relationship in different age groups, which deserves great attention.

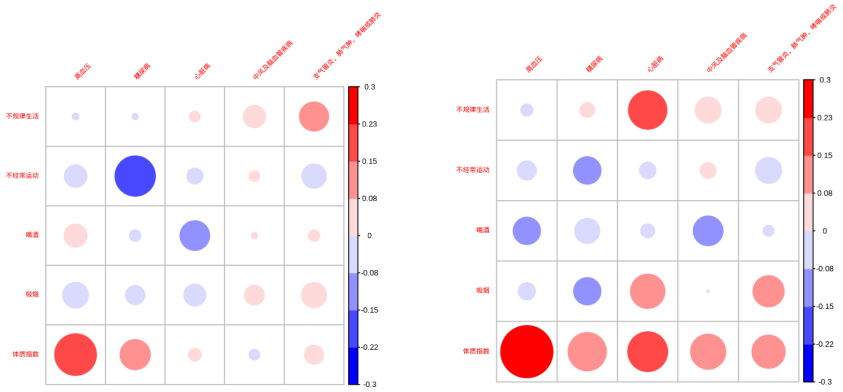
1.2 Data Information

Our data was collected by several hospitals using questionnaires and fitness tests, including 17417 people of different ages. We chose five risk factors and five non-communicable diseases as the variables. The risk factors are drinking, smoking, irregular lifestyles, lack of exercise and body mass index (BMI). The diseases are hypertension, heart disease, diabetes, apoplexy and respiratory system diseases. All those self-reported variables are binary except BMI, which is calculated and stratified in to 4 levels. After we dropped NA data, 16009 observations are used for following analysis.

1.3 Exploratory Data Analysis

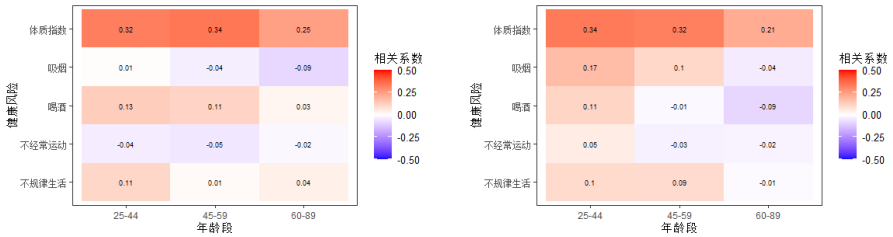
We first calculated the correlation coefficients between risk factors and diseases. Even though correlation is not equivalent to causation, it can reveal the relationships between the variables in a different aspect. To illustrate the possible

change across ageing process, we divided ages in to three groups: 25-44, 45-59 and 60-89 and did correlation analysis respectively. As all variables were discrete, we used polychoric correlations which could be implemented by R package `psych`.



(a) Male (b) Female

Figure 1: Overall correlation for males and females



(a) Male (b) Female

Figure 2: Correlations of risk factors with hypertension across age groups

Figure 1 showed the difference between males and females. Meanwhile in Figure 2, some coefficients changed heavily in ageing process. From these results, we may expect the causal relations for highly correlated variables, and such relationships may distinct between genders and age groups.

2 Method

To discover the causal relations for observational data, directed causal graphical models are frequently used. A directed causal graphical model consists of three components: a set of random variables, a set of directed edges between any

two of the variables and a joint probability distribution for all variables. The directed edges indicate that two variables are associated when other variables are fixed at some values, which provides the causal meaning to a directed graph. In other words, $X_i \to X_j$ means that the change of X_i can directly change the distribution of X_j. The joint distribution of all variables can be expressed as the multiplication of conditional probability on each variables' parents, that is

$$\Pr(X_1, \ldots, X_m) = \prod_{i=1}^{m} \Pr(X_i | \text{pa}(X_i))$$

where $\text{pa}(X_i)$ indicates the parents of X_i. Usually the directed acyclic graphs(DAGs) are used. Two assumptions should be held: (1) local Markov condition: every X_i in DAG is independent of its non-descendants conditional on its parents. (2) faithfulness condition: DAG demonstrates all conditional independence relations for the population probability distribution. More details can refer to Glymour et al. [2019]. The main algorithms to learn the DAGs are as follows.

2.1 Constrained-Based Algorithm

The constrained-based algorithms are based on the work of Verma and Pearl [1992] and first implemented as PC algorithm[Spirtes et al., 2000]. PC algorithm first constructed an undirected graph by independence and conditional independence tests. Then the edges can be directed by v-structure. This algorithm was refined by Colombo et al. [2014] to be more stable. More recent algorithms can refer to [Margaritis, 2003, Yaramakala and Margaritis, 2005]. One of the drawbacks for PC algorithm is the validity of the multiple testing. To partly conquer it, in practical, one may treat the significant level α as a tuning parameter. Another concern is that some edges may be maintained as undirected.

2.2 Score-Based Algorithm

As each DAG is connected with a unique decomposition of the joint distribution, we can fit a DAG to data by comparing the distribution. In details, each DAG is assigned to a network score reflecting its goodness of fit, and the algorithm attempts to maximize it. Several scores can be applied, such as BIC, NML[Shtar'kov, 1987], fNML, qNML[Silander et al., 2008, 2018]. Comparing with constrained-based algorithms, the score-based algorithms guarantee the edges are directed. However, the causal relations are not shown by scored-based algorithms, which may lead to poor interpretation.

2.3 Hybrid Algorithm

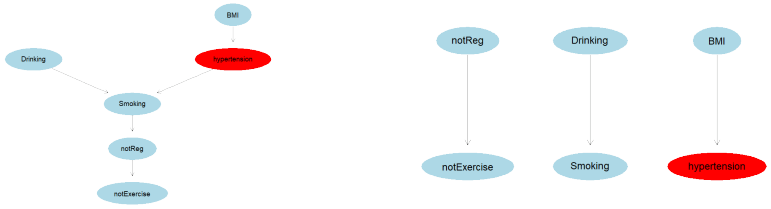
The algorithms above have their own advantages, and the improvement can be a combination of them by intuition. This can be referred as hybrid algorithms. The hybrid algorithms aim to find the undirected structure by conditional independence test, and direct the edges by score-based algorithms.

MMHC[Tsamardinos et al., 2006], RSMAX2[Scutari et al., 2014] and H2PC[Gasse et al., 2014] are frequently used hybrid algorithms.

3 Results

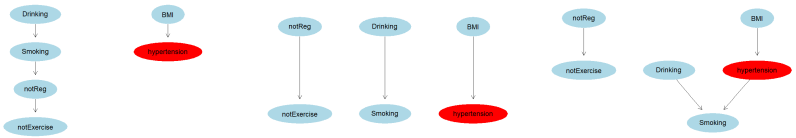
We used the hybrid algorithm to construct our causal graphical model. At the same time, the other two algorithms were applied for comparison. Confounding variables were considered in the sensitivity analysis. We used the R package `bnlearn` to construct the model for risk factors and hypertension as an example. For other diseases, one could see `main.R` for more details. Figure 3 showed the causal relations of risk factors and hypertension for males and females respectively. Different structures appeared for different genders, however, BMI was a common cause of hypertension. For males, hypertension had a direct edge to smoking, who had a direct edge to irregular lifestyles, which were distinct from females. We fit models to both genders for different age groups in Figure 4 and 3. The difference appeared across age groups, especially in the relations between risk factors. The causal relations between smoking and hypertension only appeared in Figure4(c).

A visual comparison for methods introduced in Section 2 was shown in Figure 6. Same structures were constructed for the hybrid algorithm and the score-based algorithm using qNML. As for the constrained-based algorithm, we used the PC algorithm and chose $\alpha = 0.05$ and $\alpha = 0.001$ as the significance level for conditional independence tests. In Figure 6(c), when $\alpha = 0.05$, the undirected structure was same with Figure 6(a), but the directions were partly diverse. The edge between smoking and the irregular lifestyle disappeared in Figure 6(d), due to the lower significance level. Also, undirected edges were preserved in Figure 6(c) and (d). Figure 7 showed the numerical comparison between methods. The five-fold cross-validation was applied. We fitted models to four folds of data and the rest gave the log-likelihood loss to the models. This result partly coincided with the visual comparison result as the hybrid algorithm showed no difference from the score-based algorithm. In Figure 8, we treated hypertension as a confounder when we constructed a model for risk factors and diabetes. Figure 8(a) showed no paths from BMI to diabetes while in Figure 8(b) the path existed as it went through the confounding variable.



(a) Male (b) Female

Figure 3: Risk Factors and Hypertension across all age groups



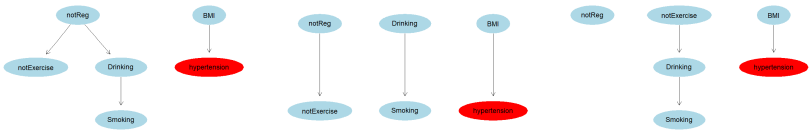
(a) 25-44 (b) 45-59 (c) 60-89

Figure 4: Causal graphs for male with different age groups

4 Discussion

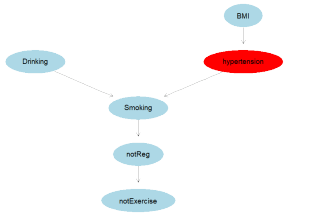
The causal relations of risk factors and disease could be discovered by graphical models. In our results, the difference appeared across different genders and age groups. BMI may be a main cause of hypertension, which was indicated by all graphs, regardless of genders and ages. Surprisingly, smoking only appears to be related with hypertension in the eldest group of males, and the direction is unexpected. This abnormal result may stem from that the incidence of hypertension may lead to smoking cessation. However, our results may be influenced heavily by confounding variables. As BMI is expected to have a causal relation with diabetes, such result only appears with the inclusion of hypertension. Therefore, involving more confounders may reverse the relations between smoking and hypertension in our model.

The causal graphical models are powerful for causal discovery. However,the existence of confounders are disturb. In future work, we may apply methods that are available for hidden confounders or ask experts for advice on effective confounders. Meanwhile, one of the restrictions of our models is that we can not tell the causal relations are positive or negative. In other words, the causal effects can not be demonstrated by our models, which is of great interest to us. To the variables which we find the causal relations, one of the future work can focus on exploring their causal effects.

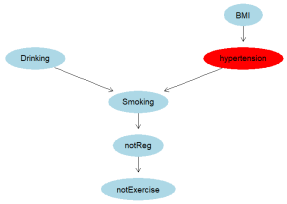


(a) 25-44 (b) 45-59 (c) 60-89

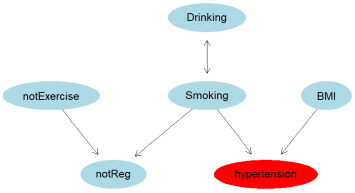
Figure 5: Causal graphs for female with different age groups



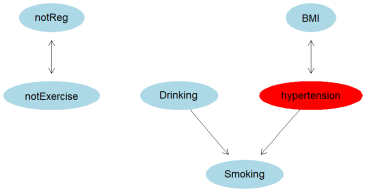
(a) Hybrid



(b) Score-based,qnml



(c) Constrained-based, PC, $\alpha = 0.05$



(d) Constrained-based, PC,$\alpha = 0.001$

Figure 6: Visual comparison for different methods

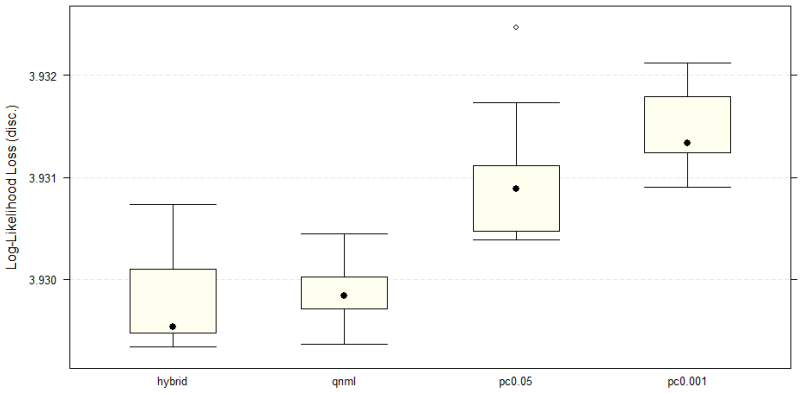
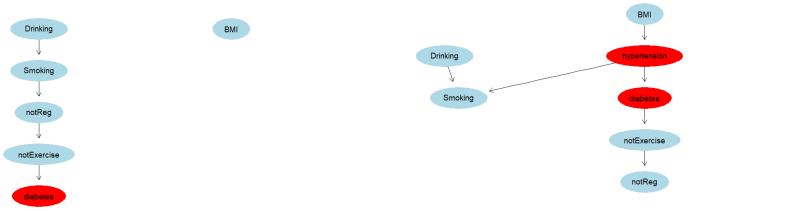


Figure 7: Log-likelihood Loss Comparison



(a) Diabetes without confounder

(b) Diabetes with Hypertension as confounder

Figure 8: Including confounders or not

References

D. Colombo, M. H. Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

M. Ezzati and E. Riboli. Behavioral and dietary risk factors for noncommunicable diseases. *New England Journal of Medicine*, 369(10):954–964, 2013.

M. Gasse, A. Aussem, and H. Elghazel. A hybrid algorithm for bayesian network structure learning with application to multi-label learning. *Expert Systems with Applications*, 41(15):6755–6772, 2014.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

D. Margaritis. Learning bayesian network model structure from data. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2003.

M. Scutari, P. Howell, D. J. Balding, and I. Mackay. Multiple quantitative trait analysis using bayesian networks. *Genetics*, 198(1):129–137, 2014.

Y. M. Shtar'kov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.

T. Silander, T. Roos, P. Kontkanen, and P. Myllymäki. Factorized normalized maximum likelihood criterion for learning bayesian network structures. In *Proceedings of the 4th European workshop on probabilistic graphical models (PGM-08)*, pages 257–272. Citeseer, 2008.

T. Silander, J. Leppä-Aho, E. Jääsaari, and T. Roos. Quotient normalized maximum likelihood criterion for learning bayesian network structures. In *International Conference on Artificial Intelligence and Statistics*, pages 948–957. PMLR, 2018.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Uncertainty in artificial intelligence*, pages 323–330. Elsevier, 1992.

WHO. Ageing and health. `https://www.who.int/news-room/fact-sheets/detail/ageing-and-health`.

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.