

Causal Structure Learning on Detection of Cardiovascular Disease (CVD) Dataset

Anushadevi Rajkumar, ID: 220210265, ECS784P – Data Analytics - 2022/23

QUESTION 1: Discuss the research area and the data set you have prepared, along with pointers to your data sources. Screen-capture part of the final version of your data set and present it here as a Figure. For example, if your data set contains 15 variables and 1,000 samples, you could present the first 10 columns and a small part of the sample size. Explain why you considered this data set to be suitable for structure learning, and what questions you expect a structure learning algorithm to answer.

ANSWER:

The research topic for this coursework is the detection of cardiovascular disease on individuals through medical examinations. Cardiovascular disease (CVD) refers to the condition that affects the heart and blood vessels which leads to heart failure, stroke, and artery diseases. [1] Research in this area is vital as understanding the factors that increase the risk of CVD leads to early detection and prevention of the disease. The dataset utilized for this is from Kaggle - [cardiovascular disease dataset](#) [2] containing a sample size of 70,000 that are reduced during the pre-processing of the data set to 4927 samples. With 3 types of input variables – factual, examination and subjective; it is feasible to draw out the cause and effects of these variables that targets in detecting the presence/absence of CVD. The dataset was pre-processed with categorizing the discrete values, replacing missing values and feature engineering.

The chosen dataset is appropriate for structural learning as the number variables involved are adequate for applying the structural learning methodologies. Through structure learning, the identification of complex relationships between variables, patterns of causal associations, mainly the combinations of different risk factors and lifestyle choices influencing the CVD are expected to be found.

| | age | gender | ap_hi | ap_lo | cholesterol | | gluc | smoke | alco | active | cardio | bmi_category |
|---|-------|--------|---------|---------|-------------------|-------------------|------|-------|------|--------|--------|---------------|
| 0 | 40-50 | Male | Normal | Normal | normal | normal | 0 | 0 | 1 | 0 | 0 | Normal |
| 1 | 50-60 | Female | Stage 2 | Stage 2 | well above normal | normal | 0 | 0 | 1 | 1 | 1 | Obese Class 1 |
| 2 | 50-60 | Female | Stage 1 | Normal | well above normal | normal | 0 | 0 | 0 | 0 | 1 | Normal |
| 3 | 40-50 | Male | Stage 2 | Stage 2 | normal | normal | 0 | 0 | 1 | 1 | 1 | Overweight |
| 4 | 40-50 | Female | Normal | Normal | normal | normal | 0 | 0 | 0 | 0 | 0 | Normal |
| 5 | 50-60 | Female | Normal | Normal | above normal | above normal | 0 | 0 | 0 | 0 | 0 | Overweight |
| 6 | 50-60 | Female | Stage 1 | Normal | well above normal | normal | 0 | 0 | 1 | 0 | 0 | Obese Class 2 |
| 7 | 60-70 | Male | Stage 1 | Stage 2 | well above normal | well above normal | 0 | 0 | 1 | 1 | 1 | Overweight |
| 8 | 40-50 | Female | Normal | Normal | normal | normal | 0 | 0 | 1 | 0 | 0 | Overweight |
| 9 | 50-60 | Female | Normal | Normal | normal | normal | 0 | 0 | 0 | 0 | 0 | Overweight |

Figure 1 - Snippet of the first 10 columns of the utilized dataset

QUESTION 2: Present your knowledge-based DAG (i.e., DAGtrue.pdf or the corresponding DAGtrue.csv graph visualised through the web editor), and briefly describe the information you have considered to produce this graph. For example, did you refer to the literature to obtain the necessary knowledge, or did you consider your own knowledge to be sufficient for this problem? If you referred to the literature to obtain additional information, provide references and very briefly describe the knowledge gained from each paper. If you did not refer to the literature, justify why you considered your own knowledge to be sufficient in determining the knowledge-based graph.

ANSWER:

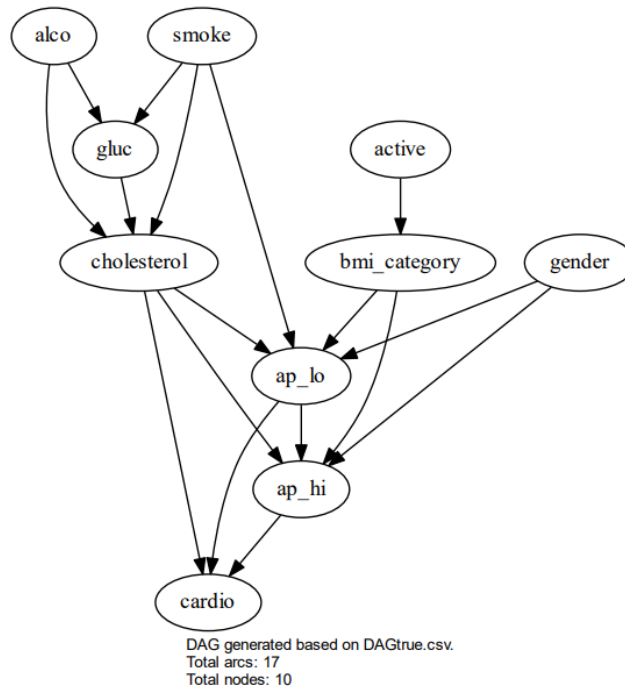


Figure 2 - Knowledge-based DAG of the CVD dataset

The above knowledge-based DAG was created based on various literature review on health journal papers. The 10 variables utilized for this graph includes:

- (1) alcohol intake (alco), smoking (smoke) and physical activity (active) being the information provided by the patients.
- (2) glucose (gluc), cholesterol and, systolic (ap_hi) and diastolic blood pressure (ap_lo)
- (3) the factual data present were the patient's gender and their Body Mass Index (bmi_category) was determined by the individual's height and weight present on the dataset.

The effects of a person's lifestyle choices have a significant impact on blood pressure, cholesterol, and glucose levels. Alcohol intake, smoking, nutrition, and physical inactivity all increase the risk of cardiovascular disease, with glucose and cholesterol levels rising as a result. [3] Despite other clinical risk factors, the relationship between physical activity and body mass index has a direct impact on blood pressure levels. According to studies, blood pressure rises with a greater BMI. [4] Studies conducted by the medical industry have shown that men have higher blood pressure levels than women, which has an impact on both systolic and diastolic pressure. [5] Systolic blood pressure rises when diastolic blood pressure is higher since this is a sign of the existence of cardiovascular illness. [6]

QUESTION 3: Complete Table Q3 below with the results you have obtained by applying each of the algorithms to your data set during Task 4. Compare your CPDAG scores produced by F1, SHD and BSF with the corresponding CPDAG scores shown in Table 3.1 (page 13) in the Bayesys manual. Specifically, are your scores mostly lower, similar, or higher compared to those shown in Table 3.1 in the manual? Why do you think this is? Is this the result you expected? Explain why.

ANSWER:

Table 1 - CPDAG scores obtained from Task 4

| Algorithm | CPDAG scores | | | Log-Likelihood (LL) score | BIC score | No. of free parameters | Structure learning elapsed time |
|-----------|--------------|--------|-------|---------------------------|------------|------------------------|---------------------------------|
| | BSF | SHD | F1 | | | | |
| HC | 0.027 | 18.500 | 0.250 | -42261.374 | -42758.167 | 81 | 0 |
| TABU | 0.027 | 18.500 | 0.250 | -42261.374 | -42758.167 | 81 | 0 |
| SaiyanH | 0.122 | 16.500 | 0.321 | -42329.881 | -42820.541 | 80 | 0 |
| MAHC | 0.034 | 18.000 | 0.231 | -42441.784 | -42822.046 | 62 | 0 |
| GES | 0.027 | 18.500 | 0.250 | -42261.374 | -42758.167 | 81 | 0 |

Comparing the results obtained to the CPDAG average scores mentioned on the Bayesys manual, it can be concluded that the scores of each algorithm are low. However, with the number of nodes, edges and sample size taken into consideration, the results obtained from the graph consisting of a total of 10 nodes, 17 arcs and a sample size of 4927 can be compared to the network of 'Sports'. From the scores under the network of sports, it is still visible that all the obtained results are lower.

The results derived from task 4 are expected as the dataset used differs in complexity compared to the network utilized on the manual. Other possible reason for this to occur could be due to the limitations present in the training dataset. It can be assumed that some latent (missing) variables could highly have influence in the dependencies of the existing variables. The possible latent variables could be family medical history, stress levels and diet as these variables can have an impact on other variables. It can be analysed from journal articles that the influence of medical history can highly increase the risk of cardiovascular disease by 40% to 75% and blood pressure (diastolic and systolic). [7] Other factors like diet and stress level can impact glucose, blood pressure and BMI levels. With latent variable being a limitation along with the complexity of the network as many nodes has an influence on one another, the results obtained are expected.

QUESTION 4: Present the CPDAG generated by HC (i.e., CPDAGlearned.pdf or the corresponding CPDAGlearned.csv graph visualised through the web editor). Highlight the three causal classes in the CPDAG. You only need to highlight one example for each causal class. If a causal class is not present in the CPDAG, explain why this might be the case.

ANSWER:

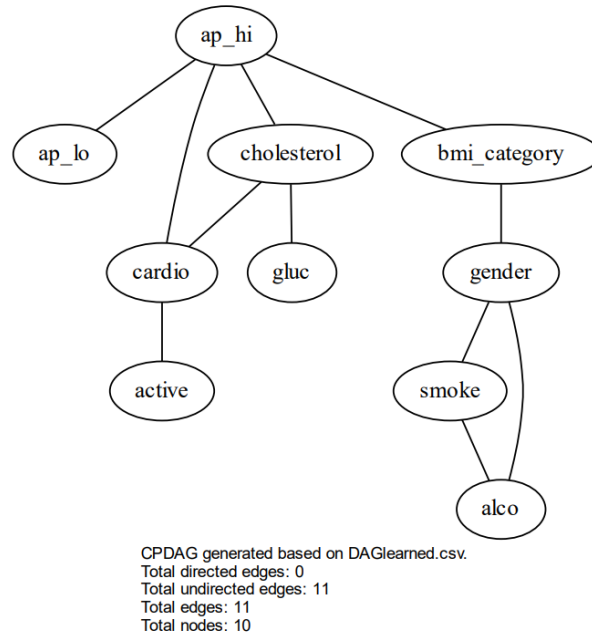


Figure 3 - CPDAG generated from HC algorithm for the utilized dataset.

From the above CPDAG generated by HC, it is noticeable that all edges are undirected. Therefore, it can be concluded that only two causal classes (causal chain and common cause) are determined. These two causal classes are represented by $X \perp Z \mid Y$, when all variables in the model are marginally independent of other given the training data. Drawing out these as two possible causal classes from the undirected edges, below is an example for each:

- (1) Causal chain: ap_hi (systolic blood pressure) \rightarrow cholesterol \rightarrow gluc (glucose); the influence of systolic pressure in cholesterol could cause an impact on cholesterol with the given dataset. The fluctuations on cholesterol can impact the glucose level.
- (2) Common cause: ap_lo (diastolic pressure) \leftarrow ap_hi (systolic pressure) \rightarrow cardio (cardiovascular disease); as shown from studies, systolic pressure can be a stronger predictor than diastolic pressure for cardiovascular disease mainly in adults over 50. Since the dataset provided for training contains majority of the results from adults over 50 it can be assumed that the ap_hi variable is a common cause influencing the other two variables. [8]

QUESTION 5: Rank the six algorithms by score, as determined by each of the three metrics specified in Table Q5. Are your rankings consistent with the rankings shown under the column “Rankings according to the Bayesys manual” in Table Q5 below? Is this the result you expected? Explain why.

ANSWER:

Table 2 - Rankings of CPDAG scores

| Rank | Rankings from obtained scores | | | Rankings according to the Bayesys manual | | |
|------|---|--|---|--|---------------------|--------------------|
| | BSF [single score] | SHD [single score] | F1 [single score] | BSF [average score] | SHD [average score] | F1 [average score] |
| 1. | SaiyanH [0.122] | HC [18.500] TABU [18.500] GES [18.500] | SaiyanH [0.321] | SaiyanH [0.559] | MAHC [50.96] | SaiyanH [0.628] |
| 2. | MAHC [0.034] | MAHC [18.000] | HC [0.250] TABU [0.250] GES [0.250] | GES [0.506] | SaiyanH [57.98] | MAHC [0.579] |
| 3. | HC [0.027] TABU [0.027] GES [0.027] | SaiyanH [16.500] | MAHC [0.231] | MAHC [0.503] | HC [62.36] | GES [0.552] |
| 4. | - | - | - | TABU [0.499] | TABU [62.63] | TABU [0.549] |
| 5. | - | - | - | HC [0.498] | GES [63.3] | HC [0.548] |

The rankings derived from the obtained scores of Task4 in comparison with the ranks from Bayesys manual are consistent in some algorithms and not in others. The ranking of BSF is somewhat similar to BSF rankings from the manual with SaiyanH [0.122] performing better under BSF and TABU [0.027] performing below average. As for SHD, it is visible that the rankings are not at all consistent with the manual, as the algorithms HC, TABU and GES performs well with a score of 18.500 than SaiyanH [16.500]. According to the F1 score, the 1st ranking is compatible with the manual for SaiyanH along with GES as the 2nd.

The inconsistency of these ranking compared to the ones from the Bayesys can be due to the incorrect dependency prediction between the variables from the given training data. This could be due to the noise present within the data where the observed connections between variables are to some unknown confounder. It is possible that the learnt causal relationship existing between the two variables are not dependent leading to causation without association.

QUESTION 6: Refer to your elapsed structure learning runtimes and compare them to the runtimes shown in Table 3.1 in the Bayesys manual. Indicate whether your results are consistent or not with the results shown in Table 3.1. Explain why.

ANSWER:

In reference to the runtime (secs) mentioned on the manual, the elapsed structure learning runtime for all algorithms are consistent comparing it to the appropriate sample size on table 3.1. It can be observed from the sports network that the sample size of 10^3 and 10^4 (since the sample size used for the training data is 4927) that it results in a runtime of 0. Comparing it to the average runtime present on the table 3.1, the runtimes are drastically lower as these indicates less complexity in the training dataset and causal relation between the variables.

QUESTION 7: Compare the BIC score, the Log-Likelihood (LL) score, and the number of free parameters generated in Task 3, against the same values produced by the five structure learning algorithms you used in Task 4. What do you understand from the difference between those three scores? Are these the results you expected? Explain why.

ANSWER:

Table 3 - Obtained BIC, Log-Likelihood scores and # of free parameters from CPDAG.

| Algorithm | Task 3 results | | | Algorithm | Task 4 results | | |
|------------------------------|----------------|----------------|-----------------|----------------|----------------|----------------|-----------------|
| | BIC score | Log-Likelihood | Free parameters | | BIC score | Log-Likelihood | Free parameters |
| Knowledge-based graph | -47070.765 | -42716.159 | 710 | HC | -42758.167 | -42261.374 | 81 |
| | | | | TABU | -42758.167 | -42261.374 | 81 |
| | | | | SaiyanH | -42820.541 | -42329.881 | 80 |
| | | | | MAHC | -42822.046 | -42441.784 | 62 |
| | | | | GES | -42758.167 | -42261.374 | 81 |

Comparing the scores obtained from the knowledge-based graph to the scores produced by the structure learning algorithms, it can be observed that the BIC score, Log-Likelihood and number of free parameters of task 4 are lower than task 3. This indicates that the algorithms used on task 4 are better in fitting the model and its dimensionality with the given BIC score. With the scores obtained for Log-Likelihood, it is also evident from the results of task 4 that it performs better with data fitting the parameters. With a greater number of free parameters on the knowledge-based graph, it can be concluded that the task 4 learning algorithms results in a less complex graph.

Given that the dataset and its variables used for training are complex, these results are expected as knowledge-based graphs are produced through one's knowledge on the domain. On the other hand, these structure learning algorithms used on task 4 are designed to optimize the model's fit and complexity avoiding overfitting or underfitting issues unlike the knowledge-based graphs. Hence to conclude, from the given dataset and its complexity it is predictable that the results from task 4 are better.

QUESTION 8: Select TWO knowledge approaches from those covered in Week 11 Lecture and Lab, i.e., any two of the following: a) Directed, b) Undirected, c) Forbidden, d) Temporal, e) Initial graph, f) Variables are relevant, and g) Target nodes. Apply each of the two approaches to the structure learning process of HC, separately (i.e., only use one knowledge approach at a time). It is up to you to decide how many constraints to specify for each approach. Then, complete Table Q8 and explain the differences in scores produced before and after incorporating knowledge. Are these the results you expected? Explain why.

ANSWER:

Table 4 - Obtained scores from before and after applying knowledge approaches.

| Knowledge approach | CPDAG scores | | | LL | BIC | Free parameters | Number of edges | Runtime |
|-----------------------------|--------------|--------|-------|------------|------------|-----------------|-----------------|---------|
| | BSF | SHD | F1 | | | | | |
| Without knowledge | 0.027 | 18.500 | 0.250 | -42261.374 | -42758.167 | 81 | 11 | 0 |
| With knowledge – directed | 0.416 | 11.500 | 0.559 | -42105.773 | -43780.150 | 273 | 17 | 0 |
| With knowledge – undirected | 0.252 | 13.500 | 0.324 | -42460.687 | -42939.081 | 78 | 12 | 0 |

The two knowledge approaches chosen are (a) directed and (b) undirected, the constraints applied for these approaches can be observed from the below table:

Table 5 – Applied Directed Knowledge Approaches and its constraints and CPDAG.

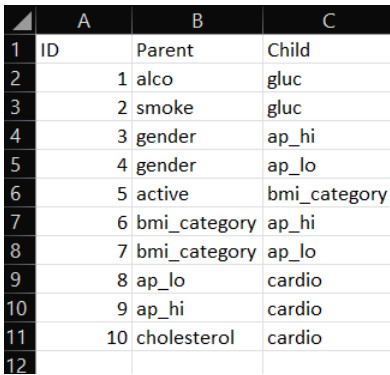
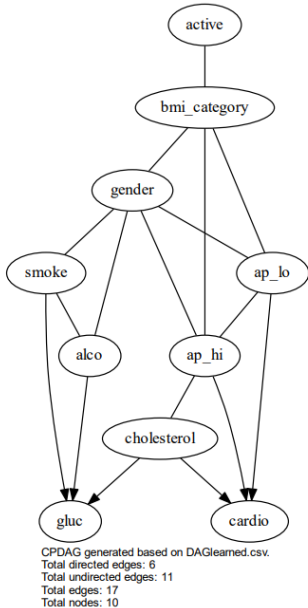
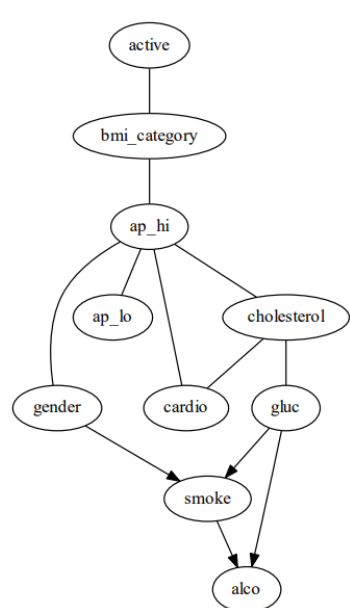
| (a) Directed Approach | | | |
|--|--|---|--|
| Directed Constraints | | CPDAG generated from HC | |
|  <p>Figure 4 - Applied directed constraints from knowledge-based graph.</p> | |  <p>Figure 5 - Generated CPDAG for directed knowledge approach.</p> | |

Table 6 – Applied undirected Knowledge Approaches and its constraints and CPDAG.

| (b) Undirected Approach | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--|----|---------------|-------------------------|---|---|---|---|----|--------|-------|---|--|---------------|--------|---|--|---------|--------|---|--|----------|-------|---|--|----------|--------------|---|--|---------|--------------|---|--|---------|------|---|--|--------|------|---|--|--|--|---|
| Undirected Constraints | | | CPDAG generated from HC | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <thead> <tr> <th></th><th>A</th><th>B</th><th>C</th></tr> </thead> <tbody> <tr> <td>1</td><td>ID</td><td>Parent</td><td>Child</td></tr> <tr> <td>2</td><td></td><td>1 cholesterol</td><td>cardio</td></tr> <tr> <td>3</td><td></td><td>2 ap_hi</td><td>cardio</td></tr> <tr> <td>4</td><td></td><td>3 gender</td><td>ap_hi</td></tr> <tr> <td>5</td><td></td><td>4 active</td><td>bmi_category</td></tr> <tr> <td>6</td><td></td><td>5 ap_hi</td><td>bmi_category</td></tr> <tr> <td>7</td><td></td><td>6 smoke</td><td>gluc</td></tr> <tr> <td>8</td><td></td><td>7 alco</td><td>gluc</td></tr> <tr> <td>9</td><td></td><td></td><td></td></tr> </tbody> </table> <p>Figure 6 - Applied undirected constraints from knowledge-based graph.</p> | | | | A | B | C | 1 | ID | Parent | Child | 2 | | 1 cholesterol | cardio | 3 | | 2 ap_hi | cardio | 4 | | 3 gender | ap_hi | 5 | | 4 active | bmi_category | 6 | | 5 ap_hi | bmi_category | 7 | | 6 smoke | gluc | 8 | | 7 alco | gluc | 9 | | | |  <p>CPDAG generated based on DAGlearned.csv. Total directed edges: 4 Total undirected edges: 8 Total edges: 12 Total nodes: 10</p> <p>Figure 7 - Generated CPDAG for undirected knowledge approach.</p> |
| | A | B | C | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | ID | Parent | Child | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | | 1 cholesterol | cardio | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | | 2 ap_hi | cardio | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | | 3 gender | ap_hi | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | | 4 active | bmi_category | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | | 5 ap_hi | bmi_category | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7 | | 6 smoke | gluc | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | | 7 alco | gluc | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

With knowledge incorporated to the graph, the scores of CPDAG, LL and BIC are improved compared to when it is not. Analyzation of the scores in terms of before and after incorporating the knowledge approaches based on the knowledge-based graph developed in task 3 are as followed:

- Directed approach – from the CPDAG scores recorded on the table 6, the score of BSF and F1 have improved while SHD has decreased in comparison with the graph without knowledge. As for the LL and BIC scores, the LL score has decreased while BIC has increased. The free parameters obtained from this approach is significantly high indicating the model's complexity and has a higher capacity in representing the dependencies between variables.
- Undirected approach – the CPDAG scores of this approach also has an improved score on BSF and F1 compared to the graph without knowledge and a lesser score on SHD. Comparing the LL and BIC scores of this to the ones without knowledge have resulted in an increase, while the free parameters are lower.

These results are expected, as adding prior knowledge enables improvement in the model by guiding it towards more accurate causal relationships. This is evident from the F1 score derived from both the knowledge-based approaches as the increase in the score indicates that there more correctly identified true positive (predicted edges) and true negatives (predicted absence). With these improvements, it can also be noticed that the constraints added has not affected the runtime as it stays at 0 secs.

REFERENCES

- [1] "World Health Organization," [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.
- [2] S. Ulianova, "Kaggle," 2019. [Online]. Available: <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>.
- [3] J. A. M. W. F. Lacombe, "The impact of physical activity and an additional behavioural risk factor on cardiovascular disease, cancer and all-cause mortality: a systematic review.," *BMC Public Health*, p. 16, 2019.
- [4] R. C. A. P. M. T. A. M. M. E. O. A. S. E. D. E. S. G. D. M. T. F. a. E. M. Francesco Landi, "Body Mass Index is Strongly Associated with Hypertension: Results from the Longevity Check-Up 7+ Study," *Nutrients*, no. PMID: 30551656, 2018.
- [5] M. Suzanne Oparil and M. Andrew P. Miller, "Gender and Blood Pressure," *The Journal of Clinical Hypertension*, vol. 7, no. 4087, p. 10, 2005.
- [6] G. P. G. Schillaci, "The dynamic relationship between systolic and diastolic blood pressure: yet another marker of vascular aging?," *Hypertens Res*, 2010.
- [7] S. C. Kolber MR, "Family history of cardiovascular disease," *Can Fam Physician*, 2014.
- [8] MORGAM Project, "Predictive Importance of Blood Pressure Characteristics With Increasing Age in Healthy Men and Women," *Hypertension*, vol. 77, no. 4, 2021.
- [9] "Minimum legal age limits," International Alliance for Responsible Drinking (IARD), [Online]. Available: <https://iard.org/science-resources/detail/Minimum-Legal-Age-Limits>. [Accessed 2022].
- [10] S. T. E. M. Emanuele NV, "Consequences of Alcohol Use in Diabetics," *Alcohol Health Res World*, no. PMID: PMC6761899, pp. 22(3):211-9, 1998.
- [11] J. Cornfield, "Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis.," no. 21, 1962.
- [12] J. K. Kruit, "HDL and LDL cholesterol significantly influence β -cell function in type 2 diabetes mellitus," *Current opinion in lipidology*, 2010.