A Bayesian network to assist the differential diagnosis of symptoms common for chronic lung disease and cardiovascular disease

No Author Given

Vrije Universiteit, Amsterdam

1 Introduction

1.1 Problem statement and research question

The present research situates itself in developing and testing the code for a Bayesian network and subsequently designing a model and running several relevant queries on it. In pursuit of the the structure and requirements given in the assignment the following paper tries to present scientifically structured research focusing on the substantial and formal requirements in the assignment. The problem statement is found in two simultaneous facts. Initially the performance of different orderings for MAP and MPE queries on growing size Bayesian networks is an active area of research. Secondarily is an own Bayesian network provided with the illustration of several self-designed queries on it. This leads to two research questions, firstly: *How do different ordering heuristics influence the performance of MAP and MPE queries for growing Bayesian network sizes?* and secondly: *To what extent do prior marginals for cardiovascular disease and chronic lung disease transform to posterior marginal distributions given a set of symptoms as evidence?*

1.2 Background

A Bayesian network is a form of probabilistic graphical models coined by Pearl[1] and later formalized by Pearl and Neapolitan.[2][6] It represents variables and their conditional dependencies through a directed a-cyclic graph. Bayesian networks are often visualized to improve the understanding of such a model, an example is given below in figure 1. The variables such as 'hear-bark' are called nodes and are connected with directed edges. These directed edges indicate a dependency between two variables, meaning that the probability of one variable being True or False depends on the truth value of the other variable. Variables may also be dependent on each other without a direct edge, as they may be connected through other nodes. If the state of a variable is known, it is called 'evidence', which can change whether other variables are dependent on one another. Variables are often written as capital letters, while evidence is often written as lower-case letters. Each variable in a Bayesian network has a conditional probability table (CPT), which contains information about the probability of a variable having a certain value, given the state of other variables. If in the example given below, 'dog-out' is True, that will increase the chance of the variable 'hear-bark' also being true. An example of this CPT is given below in figure 2.

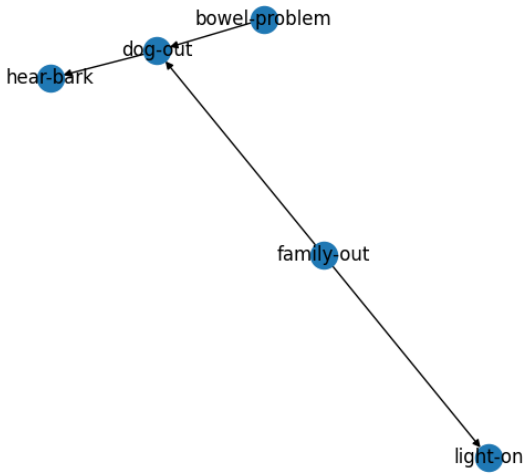

Fig. 1: An example of a directed acyclic graph from a Bayesian network

dog-out	hear-bark	p
False	False	0.7
False	True	0.3
True	False	0.01
True	True	0.99

Fig. 2: An example of a conditional probability table, p denotes the probability of a certain combination of variables

Using the conditional (in)dependencies to calculate the probability of variables being in a specific state is called inference, or a query. Such probabilities are written as Pr(A,B|C), which translates to what is the probability of A and B given C. Multiple types of queries can be made through the use of a Bayesian network. Examples are joint probabilities (Pr(A,B)), conditional probabilities (Pr(A = True|B = True)) and marginal joint probabilities. A marginal joint probability can be calculated over a joint probability by summing up all the values of the required variable probability. This represents the probability of variables having a certain value without regarding the value of another set of variables.

Furthermore, more complex queries can be instantiated, such as 'Most Probable Explanation' (MPE) and 'Maximum a Posteriori Hypothesis' (MAP). MAP is defined as the most probable assignment of some variables in a certain Bayesian network. MPE is defined similarly but computes the most likely assignment to all variables that are not in the given evidence. MAP and MPE eliminate certain variables in an order which is defined by specific heuristics, such as min-fill and

min-degree. Prior research indicates these queries are NP-hard and are far more complex to compute compared to the joint and marginal probabilities.[5] In this report, the performance of MAP and MPE queries with different heuristics will be evaluated on a set of Bayesian networks of various sizes.

Bayesian networks are used for a wide range of applications such as making predictions, anomaly detection, reasoning and diagnostics. In this report, a Bayesian network will be created to assist with the differential diagnosis of symptoms common for cardiovascular disease and chronic lung disease. These diseases are the number one and three most common causes of death respectively in the world,[10] and their symptoms are often similar. The performance of the model will be evaluated through queries and the results will be compared with logical expectations.

2 Methods: Inference Engine

In this section, an overview is given of the functionality built into the inference engine (BNReasoner) designed for this report. Various methods such as d-separation, multiple variable ordering heuristics, helper functions for queries, and MAP and MPE are described.

2.1 Helper methods

Multiplying factors Multiplying factors in the BNReasoner was implemented as a preliminary helper method, since most other methods would rely on the ability to do this. The formula took two conditional probability tables as arguments and returned the new multiplied probability table.

This was done by first constructing a new CPT of size 2^n from the variables in the two given CPTs. This new CPT was then filled with filling the rows with every possible combination of assignments of the variables. Secondly, the variables in the new CPT belonging to the first given CPT were compared to find the corresponding row and p-value in the first CPT while the variables coming from the second CPT were searched for in the second CPT yielding the p-values for their rows. This yielded two p-values for the two respective CPTs which were multiplied and assigned in the new CPT to the row corresponding to this variable instance. Notably, if a variable was present in both of the given CPTs his would result in the new CPT being a joint distribution of the two given CPTs.

Summing out and maximizing out Variables considered irrelevant for a certain query could be summed or maximized out using the corresponding functions. The intuition behind summing out can be seen as combining the case where a certain variable is True with the case where that variable is False to one merged case where the variable is either True or False and the truth value of that particular variable is considered to be irrelevant. This way, variables could be eliminated from the table. The column that corresponded to the variable that needed to be summed or maximized out was deleted. This yielded a table of pairs of rows with similar truth values for the remaining variables. In the case

of summing out, the probabilities of similar rows were added up, where as in the case of maximizing out, the row with the highest probability of a pair was selected and the other row of that pair was deleted.

2.2 *d*-separation

Implementing d-separation, an attempt was made to implement a method that determined whether it was guaranteed that two variables where independent of each other given the evidence variables in E.

Conceptually, this consisted of determining all paths between the variables and checking if any of these paths was active. This was done by checking for each path if any triple was inactive (a single inactive triple deactivates the whole path). If there were active paths, independence was not guaranteed, while if there were no active paths, independence was guaranteed.

Whether triples were active was determined by inferring the structure of each triple, i.e. V-structure, causal-chain, inverse-causal-chain and common-cause. For the former, the child had to be in the evidence for the triple to be active, while for the latter the parent or middle nodes were not allowed to be in the evidence for the triple to be active.

2.3 Variable ordering

MAP and MPE queries are very computationally expensive to compute. However, the structure of Bayesian networks can be exploited to execute such inferences faster. Variable elimination is the practice to marginalize out variables to decrease computational complexity. The order in which variables are eliminated has an impact on the resulting structure and remaining complexity of the query. The goal of variable ordering is to find the most efficient order of variables to eliminate. Two variable ordering heuristics have been implemented into the inference engine to compare alongside a random ordering. The 'min-fill' heuristic calculates and orders each variable based on how many edges would need to be added upon its removal, which is based on the intuition that adding new edges introduces unwanted complexity. The 'min-degree' heuristic orders variables based on how many neighbors it currently has, this heuristic is based on the intuition that deleting variables with fewer neighbors introduces less complexity compared to variables with many neighbors. Lonely nodes are deleted first.

2.4 Network pruning

The process of network pruning is an iterative one, which is repeated until the network can not undergo any further pruning. The network is pruned based on a given query and evidence variables. During each iteration, three actions are performed. First, any leaf nodes which are not part of the query or evidence variables are removed from the network. The other two actions are performed together. For each node given in the evidence, the edges from that node to its children are removed. Since the probabilities in the child's CPT are then independent of its parent, the CPT is updated such that it reflects this change.

2.5 Marginal joint distribution

A method for computing a marginal joint distribution was implemented. This function was dependent on the implementation of getting a factor product and required the arguments Q and E where the former was a set of variables for which the joint probability distribution was wanted, and the latter was a set of evidence variables which needed to be marginalized out. The concept of marginalization happened by summing the p-values over the rows where the variables in Q were the same, but the variables in E had to marginalized out. This was implemented by getting the CPTs for the variables in Q and subsequently multiplying all the requested CPTs into one big CPT. After that, the rows containing the variables with truth assignments corresponding to the evidence were dropped, resulting in a CPT with the variables in Q and not in E as columns and an additional column p. Now the rows which had equal values for their respective Qs were grouped, while in this grouping process the p-values of the grouped rows were summed. This provided an efficient implementation of marginalizing out variables.

2.6 MPE and MAP

Most **p**robable **e**xplanation and **m**aximum **a** **p**osteriori estimate queries are two very similar queries for retrieving a likely instantiation given evidence. Noting their similarity, this section will explain MPE, and point out the differences with MAP.

MPE starts by pruning the network, as is described in section 2.4. This pruned network is then conditioned on the evidence by removing all rows from all CPTs that contradict the evidence. The remaining variables are ordered by one of the ordering functions introduced in section 2.3. An iterative process follows, where for each variable, the product of all CPTs in which the variable occurs is calculated. The variable is maxed out over the new CPT, and all the CPTs used to create the product are replaced by the new CPT. In addition to the evidence required by MPE, MAP also requires a set of query variables. This leads to changes in three parts. First, the pruning now takes into consideration these query variables. Second, the variable elimination order is altered such that the query variables occur at the end. And last, maxing out does not occur in each iteration. If the variable, that would be maxed out in MPE, occurs in the set of query variables, it is still maxed out. Otherwise, the variable is summed out over the CPT.[5]

3 Performance

3.1 Setup

The performance of the MAP and MPE queries with different heuristics was measured on automatically generated random Bayesian networks of various sizes. A function was written to do this. The sizes of these networks ranged from 10 to 70 variables, with increments of 10 variables between networks. The number of edges between variables ranged from 19 to 363 edges. Because of the random nature of the network generation, results were averaged over 10 randomly generated networks for each size. Performance of the different heuristics was measured

as time in seconds to completion. The results for MAP and MPE are shown in figure 3.

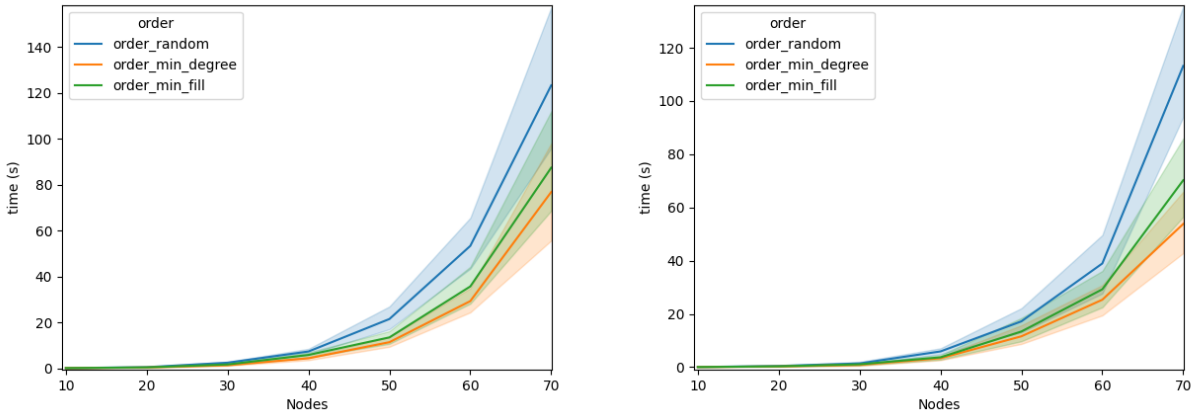

Fig. 3: Performance of various heuristics, measured in terms of computation time needed to answer a MAP query (left) or MPE query (right)

3.2 Results and statistics

Both performance plots show that small networks do not discriminate between different methods of ordering. The results show that computation time grows logarithmic with increasing network size. A trend of the worst performance for random ordering and best performance for the min-fill heuristics can be read from both plots. The differences seem to enlarge with increasing number of nodes, although not all differences tested significant. This could possibly be explained by the fact that the group size of 10 networks per group was rather small. A one-way ANOVA test was used to test for significant differences between groups. P smaller than 0.05 was considered statistically significant.

For MAP, the networks of 30, 50 and 60 nodes performed significantly better for the two heuristics as compared to random ordering. For 40 and 70 nodes, only min degree was significantly faster than random ordering. For networks of 10 or 20 nodes, no significant differences were found. For MPA, no no significant differences were found for network sizes up to 30 nodes, as well as for 50 and 60 nodes networks. For networks of 40 and 70 nodes, both heuristics differed significantly from random ordering. Overall, the plots shown a clear trend of better performance for the min degree and min fill heuristic, although not always significant.

3.3 Discussion and conclusion

Several noteworthy patterns emerged when performing analysis on the aforementioned results. A few remarks in different categories can be stated.

First of all, results on both MAP and MPE queries show that random ordering performs worst in terms of computation times. Min degree performs best, leaving the performance of min degree somewhere in between. The worst performance of random ordering is obvious and could be explained by the fact that it does not exploit the benefits of a smart elimination order, making the computations less efficient. This is in accordance with the literature on MAP and MPE queries, and several aspects of their performance seem to indeed indicate the computational benefit MPE has over MAP.The logarithmic performance of both algorithms with regard to the number of nodes present in the network and the different heuristic display performance differences according to expectation.

For MPE, the difference between min-degree and min-fill was notably more pronounced than for MAP. This observation is noteworthy considering the expected larger number of nodes that would need to be ordered for a MAP query. It is hypothesized that the higher degree of uncertainty in the intermediary variables in MAP creates computational complexity not directly related to the heuristics, causing their difference to be less pronounced. Their simultaneous difference to random-ordering was more pronounced in MAP queries. While this is initially expected as the number of nodes in the intermediary nodes is larger, this argument would create a dichotomy with the previous hypothesis that the complexity of the intermediary nodes decreases the clear difference between the heuristics in MAP. = A sufficient hypothesis for the presently obtained results must account for the increased difference between the heuristics in MPE and the decreased difference with random ordering in MPE. This led the present research to attempt to harmonize these opposing phenomena by assuming the unique role a random ordering takes in the queries, while min-degree and min-fill are indirectly related to each other. Future research might further investigate the apparent contradiction of the pronounced difference of the performance of the different heuristics under different query types.

4 Modeling use-case

4.1 Setup and explanation use-case

Cardiovascular disease and chronic lung disease are the number one and three most common causes of death in the world respectively[10]. These diseases have many nonspecific symptoms in common such as tightness in chest, shortness of breath and metabolic syndrome, this makes differential diagnosis based on these symptoms very difficult. Several items of research have been produced in the area of using Bayesian networks for medical diagnosis.[7][4][8][3][9] This report introduces a Bayesian network to assist in the diagnosis of these diseases and includes common risk factors such as smoking, age and obesity to attempt to improve accuracy.

The Bayesian network consists of three root nodes, which are the risk factors 'Smoker', 'Older than 65' and 'Obesity'. These are introduced to have an effect

on the probability of certain symptoms being present. The probability values for the CPTs of these variables were established by consulting the several internet sources and basic reasoning.

Obesity	p	older_than_65	p	Smoker	p
False	0.819	False	0.8	False	0.81
True	0.181	True	0.2	True	0.19

Fig. 4: Conditional probability tables of the three root nodes 'Smoker', 'Older than 65' and 'Obesity'

The three root nodes shown in figure 4 directly impact the probability of various symptoms such as 'High blood pressure', 'Left ventricular hypertrophy', 'Metabolic syndrome', 'Decreased respiratory muscle endurance', 'Shortness of breath' and 'Tightness in chest'. The CPTs for these variables are constructed in such a way that having a combination of symptoms which is slightly more specific to either chronic lung disease or cardiovascular disease increases the probability of that disease being true. These symptoms are connected to more medical conditions such as 'Pulmonary fibrosis', 'Atherosclerosis' and 'Decreased mental alertness', which are ultimately connected to the two diseases 'Chronic lung disease' and 'Cardiovascular disease'. The CPTs for all of the variables can be found in the appendix. The full directed acyclic graph can be found in the figure 5 in the appendix.

As stated in the introduction, the performance of the Bayesian network will be evaluated by using the implemented inference engine to perform queries, and comparing the results to logical reasoning. The main use case of the Bayesian network introduced in this paper is to distinguish between chronic lung disease and cardiovascular disease diagnoses based on risk factors and symptoms present. For this reason, the queries used to test the Bayesian network reflect this main use case. The following queries will be executed:

Query 1a and 1b: Two a-priori marginal queries over the respective variables of the set *Chronic lung disease, Cardiovascular disease* i.e. explicating the a-priori chances of having a chronic lung disease and a cardiovascular disease.

Query 2a and 2b: Two posterior marginal query over the respective variables in the set *Chronic lung disease, Cardiovascular disease* and for both the evidence set *Tightnessinchest, Highbloodpressure, Leftventricularhypertrophy* mapping to the respective values of *True, True, True.*

Query 3: A MAP query for *Chronic lung disease, Cardiovascular disease, Smoker, Obesity* and the evidence set *Tightness in chest, High blood pressure, Left ventricular hypertrophy* mapping to the respective values of *True, True, True.*

Query 4: A MPE query where *Chronic lung disease* is in the evidence set and everything else is in the variable set, so noting what the most likely symptoms are given chronic lung disease i.e. *P(Symptoms | Chronic lung disease)*.

Query 5: Another MPE query where *Cardiovascular disease=True, Obesity=False* are in the evidence set and everything else is in the variable set so noting what the most likely symptoms are given chronic lung disease i.e. *P(Symptoms | Cardiovascular disease, ¬Obesity*). Notable for both of the latter queries is the most probable explanation for symptoms being present given the presence of one of the two types of diseases.

4.2 Results

The following section introduces the results of the forementioned queries on the conceptual model.

Table 1: Query 1a & 1b

Chronic lung disease	**p**
True	0.17618
False	0.82382

Cardiovascular disease	**p**
True	0.247944
False	0.752056

Table 2: Query 2a & 2b

Chronic lung disease	**p**
True	0.056025
False	0.105137

Cardiovascular disease	**p**
True	0.090991
False	0.040342

Table 3: Query 3, 4 and 5

MAP	**MPE**	**2nd MPE**
Smoker	**Smoker**	**Leftventricularhypertrophy**
True	True	False

4.3 Analysis and discussion

Analyzing the forementioned queries on the conceptual Bayesian network in light of the second research question leads to the conclusion that the prior marginals transform notably to posterior marginals given the evidence. The following paragraph discusses these findings more thorough.

First, one notes the a-priori unlikeliness of both chronic lung disease and cardiovascular disease. Without any evidence, the a-priori marginal chance of having any of these two is very low. Nevertheless, for a prior, a 17% and 24% chances for the respective diseases are high. Evidently, is the percentage of the population that does have these diseases not this high. The marginal prior probabilities were not given to the model but retrieved from the marginal distribution query iterating over the chain of variables leading to these variables. These high priors are probably caused by model design and variable interaction. This paper also notes that such high priors might not be far-fetched in the context in which

the model was designed. The a-priori chances of someone being investigated with the model are much higher to have any two of the diseases since there is already have the hidden given that they are investigated. A neurologist much more extensively checks the symptoms of the people in his office than does a normal doctor, since people being in his office makes his prior already different. In the population, 5% may have a disease, while 20% of the population in his office may have it. We hypothesize a similar dynamic causing the high priors in the network by the aforementioned hidden given.

Second, is the particularity of the a-posteriori queries noted. Given the evidence of tightness in the chest, high blood pressure and left ventricular hypertrophy, the chances of having either chronic lung disease or cardiovascular disease increase. This increase is notable, with the risk of having a chronic lung disease becoming approximately $\frac{1}{3}$ and the risk of having a cardiovascular disease becoming approximately $\frac{2}{3}$. So given these few items of evidence, the risk for either of these diseases increases strongly.

Finally, one notes that when only given the evidence of chronic lung disease to either the MAP or the MPE query, the network thinks the most probable instantiation of 'symptoms' is that the person is a smoker. Subject for discussion might be the societal sensitivity of the finding of the present model that the most probable instantiation for any of the two diseases is that the person is a smoker. The probability of the person having a disease given being a smoker is much easier to explain than the most probable instance of variables that if the person has a disease, the person is a smoker. This furthermore places emphasis on the danger of generalization of the most probable instantiation. For a single case this might be the most probable instantiation, but when approaching a larger number of cases the most probable instantiation strongly loses its statistical value, since the likeliness of the most probable instantiation is hidden behind the fact that it is the most probable instantiation. This percentage might be small even though this is the most likely instantiation and over larger n this percentage becomes even smaller.

4.4 Conclusion

The research in pursuit of the second research question noted that even though the prior is already strong this prior is considerably increased when noting the posterior given the evidence. Analyzing and discussing the performed queries on the particular use-case, this research noted that priors can be contextual and large priors need not always be wrong, that evidence increases the likelihood of disease as expected and that a most likely instantiation might not be generalizable to larger populations considering the small probability accompanying this most likely instantiation. This instantiation might be unexpected.

References

[1] Judea Pearl. "Bayesian netwcrks: A model cf self-activated memory for evidential reasoning". In: *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*. 1985, pp. 15–17.

[2] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. en. Google-Books-ID: AvNID7LyMusC. Morgan Kaufmann, Sept. 1988. ISBN: 978-1-55860-479-7.

[3] Agnieszka Onisko, Marek J. Druzdzel, and Hanna Wasyluk. "A Bayesian network model for diagnosis of liver disorders". In: *Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering*. Vol. 2. Citeseer, 1999, pp. 842–846.

[4] Daniel-Ioan Curiac et al. "Bayesian network model for diagnosis of psychiatric diseases". In: *Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces*. ISSN: 1330-1012. June 2009, pp. 61–66. DOI: 10.1109/ITI.2009.5196055.

[5] Johan Kwisthout. "Most probable explanations in Bayesian networks: Complexity and tractability". en. In: *International Journal of Approximate Reasoning*. Handling Incomplete and Fuzzy Information in Data Analysis and Decision Processes 52.9 (Dec. 2011), pp. 1452–1469. ISSN: 0888-613X. DOI: 10.1016/j.ijar.2011.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0888613X11001095 (visited on 12/21/2021).

[6] Richard E. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. en. Google-Books-ID: 7X5KLwEACAAJ. CreateSpace Independent Publishing Platform, June 2012. ISBN: 978-1-4774-5254-7.

[7] Gustavo Arroyo-Figueroa and Luis Enrique Sucar. "A temporal Bayesian network for diagnosis and prediction". In: *arXiv preprint arXiv:1301.6675* (2013).

[8] Mukesh Kumari, Rajan Vohra, and Anshul Arora. "Prediction of diabetes using Bayesian network". In: (2014). Publisher: Citeseer.

[9] Flávio Luiz Seixas et al. "A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer s disease and mild cognitive impairment". In: *Computers in biology and medicine* 51 (2014). Publisher: Elsevier, pp. 140–158.

[10] National Center for Health Statistics and Melonie Heron. *Deaths: Leading Causes for 2018*. en. Tech. rep. National Center for Health Statistics, May 2021. DOI: 10.15620/cdc:104186. URL: https://stacks.cdc.gov/view/cdc/104186 (visited on 12/22/2021).

Appendix

Graph for use-case

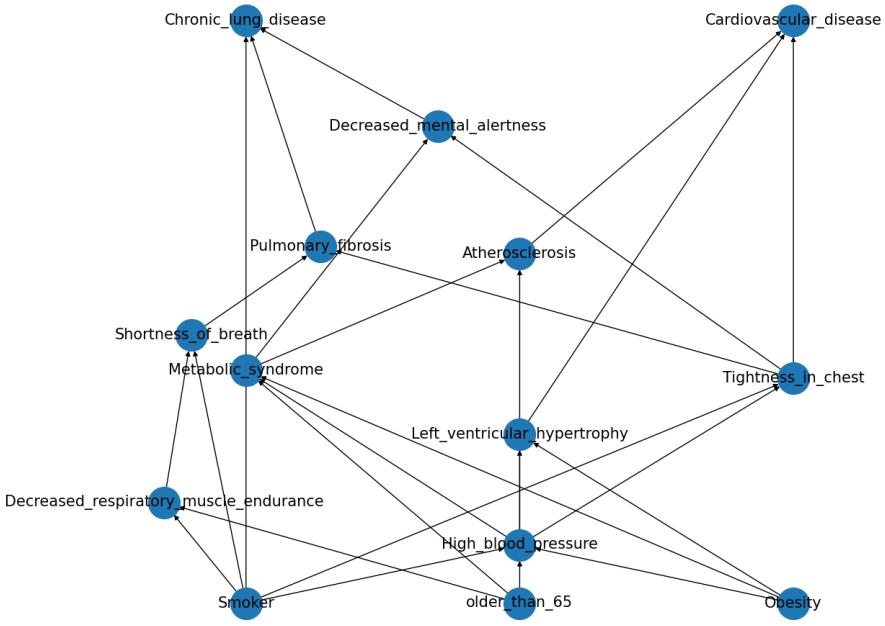

Fig. 5: A directed acyclic graph of a Bayesian network made to assist with the diagnosis of chronic lung disease and cardiovascular disease

CPTs for use-case

older_than_65

	older_than_65	p
0	False	0.8
1	True	0.2

Obesity

	Obesity	p
0	False	0.819
1	True	0.181

Smoker

	Smoker	p
0	False	0.81
1	True	0.19

Shortness of breath

	Decreased_respiratory_muscle_endurance	Smoker	Shortness_of_breath	p
0	False	False	False	0.9
1	False	False	True	0.1
2	False	True	False	0.3
3	False	True	True	0.7
4	True	False	False	0.5
5	True	False	True	0.5
6	True	True	False	0.2
7	True	True	True	0.8

High blood pressure

	older_than_65	Smoker	Obesity	High_blood_pressure	p
0	False	False	False	False	0.95
1	False	False	False	True	0.05
2	False	False	True	False	0.30
3	False	False	True	True	0.70
4	False	True	False	False	0.40
5	False	True	False	True	0.60
6	False	True	True	False	0.20
7	False	True	True	True	0.80
8	True	False	False	False	0.35
9	True	False	False	True	0.65
10	True	False	True	False	0.25
11	True	False	True	True	0.75
12	True	True	False	False	0.20
13	True	True	False	True	0.80
14	True	True	True	False	0.05
15	True	True	True	True	0.95

Decreased respiratory muscle endurance

	Smoker	older_than_65	Decreased_respiratory_muscle_endurance	p
0	False	False	False	0.99
1	False	False	True	0.01
2	False	True	False	0.75
3	False	True	True	0.25
4	True	False	False	0.60
5	True	False	True	0.40
6	True	True	False	0.20
7	True	True	True	0.80

Metabolic syndrome

	High_blood_pressure	Obesity	older_than_65	Metabolic_syndrome	p
0	False	False	False	False	0.05
1	False	False	False	True	0.95
2	False	False	True	False	0.85
3	False	False	True	True	0.15
4	False	True	False	False	0.70
5	False	True	False	True	0.30
6	False	True	True	False	0.50
7	False	True	True	True	0.50
8	True	False	False	False	0.75
9	True	False	False	True	0.25
10	True	False	True	False	0.60
11	True	False	True	True	0.40
12	True	True	False	False	0.50
13	True	True	False	True	0.50
14	True	True	True	False	0.20
15	True	True	True	True	0.80

Left ventricular hypertrophy

	Obesity	High_blood_pressure	Left_ventricular_hypertrophy	p
0	False	False	False	0.97
1	False	False	True	0.03
2	False	True	False	0.85
3	False	True	True	0.15
4	True	False	False	0.65
5	True	False	True	0.35
6	True	True	False	0.25
7	True	True	True	0.75

Tightness in chest

	Smoker	High_blood_pressure	Tightness_in_chest	p
0	False	False	False	0.90
1	False	False	True	0.10
2	False	True	False	0.70
3	False	True	True	0.30
4	True	False	False	0.50
5	True	False	True	0.50
6	True	True	False	0.25
7	True	True	True	0.75

Atherosclerosis

	High_blood_pressure	Metabolic_syndrome	Atherosclerosis	p
0	False	False	False	0.93
1	False	False	True	0.07
2	False	True	False	0.65
3	False	True	True	0.35
4	True	False	False	0.55
5	True	False	True	0.45
6	True	True	False	0.17
7	True	True	True	0.83

Pulmonary fibrosis

	Tightness_in_chest	Shortness_of_breath	Pulmonary_fibrosis	p
0	False	False	False	0.99
1	False	False	True	0.01
2	False	True	False	0.80
3	False	True	True	0.20
4	True	False	False	0.90
5	True	False	True	0.10
6	True	True	False	0.60
7	True	True	True	0.40

Decreased mental alertness

	Tightness_in_chest	Metabolic_syndrome	Decreased_mental_alertness	p
0	False	False	False	0.85
1	False	False	True	0.15
2	False	True	False	0.60
3	False	True	True	0.40
4	True	False	False	0.75
5	True	False	True	0.25
6	True	True	False	0.35
7	True	True	True	0.65

Cardiovascular disease

	Left_ventricular_hypertrophy	Atherosclerosis	Cardiovascular_disease	p
0	False	False	False	0.95
1	False	False	True	0.05
2	False	True	False	0.65
3	False	True	True	0.35
4	True	False	False	0.50
5	True	False	True	0.50
6	True	True	False	0.20
7	True	True	True	0.80

Chronic lung disease

	Smoker	Decreased_mental_alertness	Pulmonary_fibrosis	Chronic_lung_disease	p
0	False	False	False	False	0.97
1	False	False	False	True	0.03
2	False	False	True	False	0.60
3	False	False	True	True	0.40
4	False	True	False	False	0.80
5	False	True	False	True	0.20
6	False	True	True	False	0.30
7	False	True	True	True	0.70
8	True	False	False	False	0.75
9	True	False	False	True	0.25
10	True	False	True	False	0.35
11	True	False	True	True	0.65
12	True	True	False	False	0.55
13	True	True	False	True	0.45
14	True	True	True	False	0.05
15	True	True	True	True	0.95