ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

A Probabilistic Model To Determine Potential Cardiovascular Diseases Given Individual Lifestyles

Amna Shifia Nisafani, Arif Wibisono, Adi Cipta Airlangga
Institut Teknologi Sepuluh Nopember
Jalan Raya ITS, Sukolilo, Surabaya, Indonesia 60111
Email: wibisono@is.its.ac.id

Abstract

Heart diseases introduces a great number of fatalities worldwide. The vast majority of heart diseases are due to unhealthy lifestyle. Unfortunately, these lifestyles are widely unknown until the disease appears. This paper aims at developing a probabilistic model that can help individuals to early detect what type of cardiovascular disease that an individual may have if he/she maintain his/her current lifestyle. We identify factors that cause cardiovascular diseases as well as their intertwined relationships by interviewing cardiovascular experts. Subsequently, we construct Bayesian Network (BN) model based on these factors, and conduct sensitivity analysis. From our study, we obtain 14 lifestyles of cardiac and their relationships. Furthermore, there are 15 nodes of BN model to predicts cardiovascular diseases. In addition, based on our sensitivity analysis, we figure that Congenital Diseases, Type of Exercise, Body Mass Index, and Age are the most important factors contributing to cardiovascular diseases.

Keywords*: Bayesian Networks, Cardiovascular Diseases, Recommendation System*

1. Introduction

In 2013, Cardiovascular diseases (CVD) is the leading cause of mortality worldwide which constitutes of 43% for male deaths and 49,7% for female deaths [1]. Furthermore, CVD introduces high mortality in UK annually. It counts more than 160,000 individuals passed away every year [2]. According to the recent report, it is mentioned that CVD kills one out of three people in US [3]. Hence, it is important to do anything possible to reduce CVD as the number one killer globally.

There are two factors that contributes to the CVD making namely modifiable factors and non-modifiable factors [4]. Modifiable factors are any factor that able to trigger CVD because of unhealthy lifestyle such as smoking, overweight, bad nutrition, and lack of exercises [5] [6] [7][8]. While non-modifiable factors are factors that naturally inherited from both parents and are unable to alter. To this point, it is pivotal to promote healthy lifestyle in order avoid modifiable factors. The problem with this lifestyle is that many people do not realize that they are at risk. Many people are start to aware the effects of these unhealthy lifestyle after they suffer from CVD.

This research objective is to develop a probabilistic model of CVD. This model can help people to predict their possibility of having CVD in their later life given their lifestyles. There are several previous researches that have developed machine learning model in predicting CVD. Methaila et al. compared several data mining methods for forecast heart diseases. They use datasets from a Saudi Arabian hospital [9]. This paper predicts the likelihood of a patient to embrace CVD given his/her medical records such as age, blood pressure, chest pain type, and blood sugar[9]. This research is not designed to satisfy common users with lack of medical knowledge [9]. Therefore, it is impossible for them to use the model in predicting the CVD existences. Vijayashree et al. reviewed several data mining and hybrid intelligent techniques to assist medical doctors in diagnosing CVD [10]. Similar to Methaila et al. these intelligent models are only suitable for skillful physicians not commoners. Another researcher, Florence et al. introduced the combination of decision tree and artificial neural networks to predict heart attack of cardio patients. They employ health medical records to construct the model. From this point, we can infer that many researches exploited patients' medical records. Hence, the models are intended to support medical practitioners. The problem with these approaches is that common people do not have access in gaining self-diagnosing capabilities. As a result, CVD prevention approaches are very challenging to implement. Henceforward, our research aims at filling the gap that previous intelligent models do not cover.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

2. Research Method

This study consists of two phases, identifying heart disease causes and constructing probabilistic model (See Figure 1.).

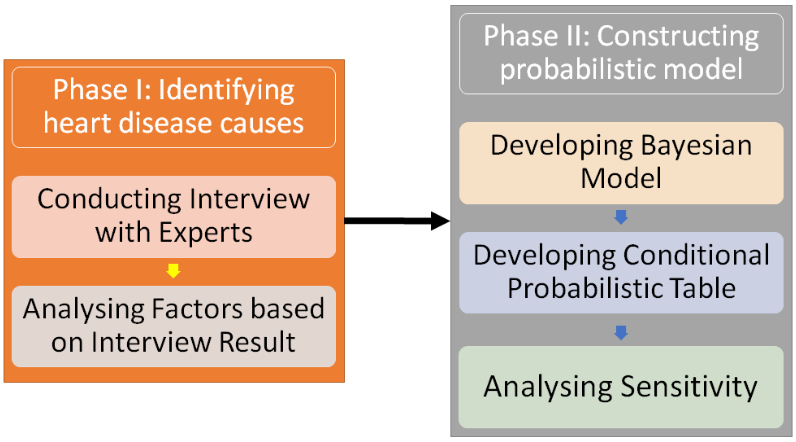
Phase I: Identifying heart disease causes
Conducting Interview with Experts
Analysing Factors based on Interview Result
Phase II: Constructing probabilistic model
Developing Bayesian Model
Developing Conditional Probabilistic Table
Analysing Sensitivity

Figure 1. Research Methodology

For the first phase, we employ qualitative method i.e. expert judgment to determine factors causing cardiovascular diseases in terms of patients' lifestyle. In addition, we identify the relationship between these factors. Before we conduct interview session with cardiovascular experts, we develop an interview form which consists of several questions as described in Table 1.

Table 1. Interview Questions

No	Questions
1	What types of cardiovascular disease that mostly attack patients
2	For each type of cardiovascular disease, what are the causes?
3	For each causes, what are their rank/weight based on their contribution to heart disease?
4	Please explain the intertwined among these causes
5	For each causes, what kind of states that usually occur based on patients profile

Based on our interview, we analyse these factors using fishbone diagram. The diagram will be used as a ground to construct probabilistic model in the following phase. We develop probabilistic model using Bayesian Network. Subsequently, we construct conditional probability table (CPT). We develop a questionnaire to capture patients profile and use it to develop CPT. The questionnaire will measure the states of each factor identified in the previous phase. Ultimately, we conduct sensitivity analysis to investigate the most contributed factors in determining cardiovascular type.

3. Results and Analysis

This section explains the result of our research which divided into four sub sections. The first sub section is Factor Analysis followed by our proposed Bayesian Model, Conditional Probabilistic Table, and Sensitivity Analysis respectively.

3.1. Factor Analysis

According to our interview, we obtain three types of cardiovascular disease that commonly attack patient, Coronary Heart Disease, Congestive Heart Failure, and Arrhythmia. In general, there are four factors causing heart disease namely patient current condition, habitual activity, diet, and congenital disease. The description for each factors can be seen in Table 2.

Table 2. Identified Factors

No	Factors	Description
1	Patient current condition	This factor represents the current condition of patient in terms of age, body mass index (BMI), heart rate/minute

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

No	Factors	Description
2	Habitual activities	This factor portrays bad habits such as smoking, lack of exercises and sleeping habits (for coronary heart disease), and working environment and fatigue (for congestive heart failure)
3	Diet	This factor describes diet habits exhibited by cardiac such as inappropriate meal portion and irregular meal frequency (for coronary heart disease), and malnutrition
4	Congenital disease	This factor captures types of congenital disease posed by cardiac such as diabetes, hypertension, hereditary heart disease, and cholesterol disorders (for coronary heart disease)

The fishbone diagram for each type of cardiovascular disease is illustrated in Figure 2 and 3 consecutively. For Arrhythmia, the major factor causing Arrhythmia is Congenital Diseases.

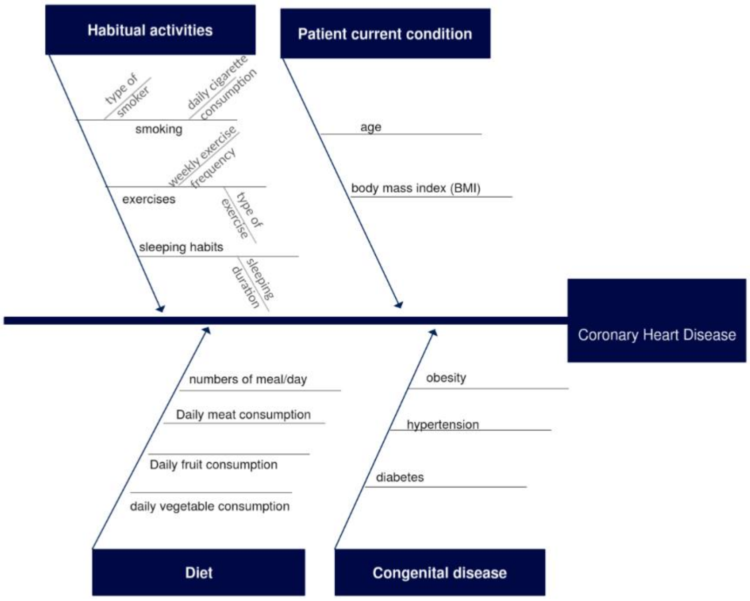

Figure 2. Fishbone diagram for coronary heart disease

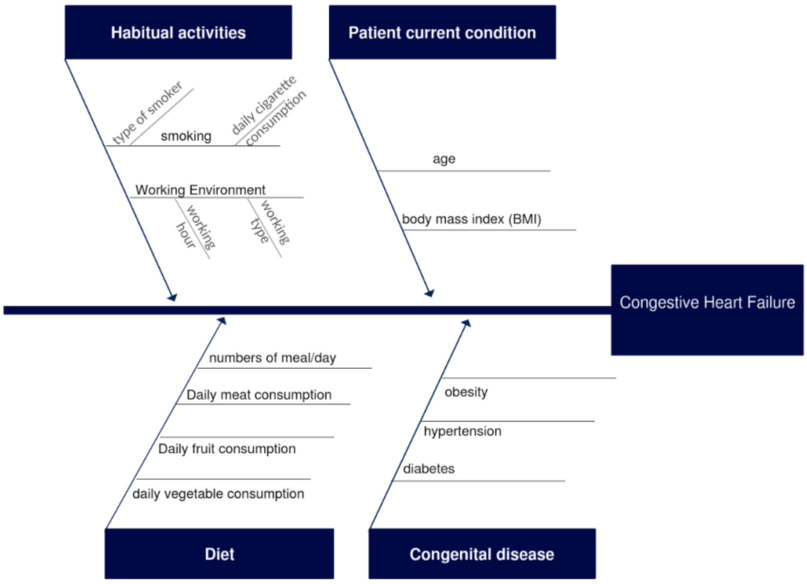

Figure 3. Fishbone diagram for congestive heart disease

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

3.2. Bayesian Model

For developing Bayesian Model, we examine states for each factor as a base for constructing questionnaire to capture cardiac profile. These states are derived from interview result and presented in Table 3.

Table 3. States of Factors

No	Factors	States
1	Type of smoker	Active Smoker
		Passive Smoker
2	Daily cigarette consumption	between 1 and 3 bars
		between 4 and 6 bars
		More than 6 bars
3	Type of exercise	Aerobic
		Anaerobic
4	Weekly exercise duration	Less than 6 hours
		Between 6 and 10 hours
		More than 10 hours
5	Sleeping duration	Less than 6 hours/day
		Between 6 and 8 hours/day
		More than 8 hours/day
6	Working hour	Less than 6 hours
		Between 6 and 10 hours
		More than 10 hours
7	Working type	Heavy labor
		Desk Job
8	Age	Less than 18 years
		Between 18 and 30 years
		Between 30 and 60 years
		More than 60 years
9	BMI	Less than 19
		Between 19 and 25
		More than 25
10	Number of meal/day	Less than 3 times/day
		3 times/day
		More than 3 times/day
11	Daily meat consumption	Do not consume
		Less than 100 grams
		Between 100 and 300 grams
		More than 300 grams
12	Daily vegetable consumption	Less than 300 grams
		Between 300 and 500 grams
		More than 500 grams
13	Daily fruit consumption	Do not consume
		Less than 100 grams
		Between 100 and 200 grams
		More than 200 grams
14	Congenital Disease	Obesity
		Diabetes
		Hypertension

We develop causative diagram by using something so-called “test approach”. In the “test approach”, an event in the target node (or root) is determined by two or more its independent children. Every child then may be split into two or more children. The “test approach” itself is inspired by Naïve Bayes, but in a larger hierarchical scale. Nevertheless, the “test approach” structure is distinct to a Naïve Bayes structure (see Figure 4).

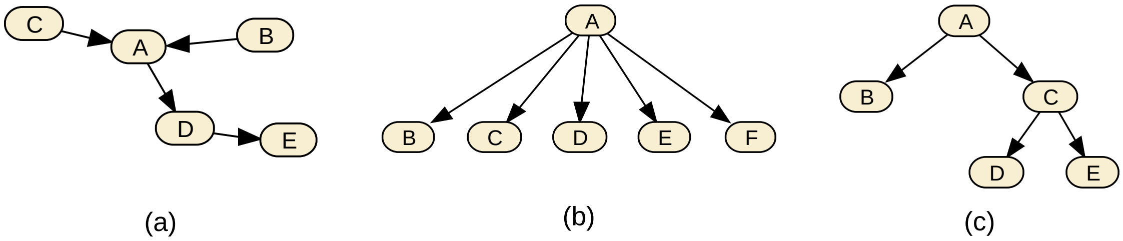

Figure 4. Various type of Bayesian Network, Focus Node A as the target node
(a) Non Naïve / Non-test approach (b) Naïve Bayes (c) Test Approach

By having this "test approach", we could reduce computational complexity of the BN model. In the BN model we employ 15 nodes as seen in Figure 5.

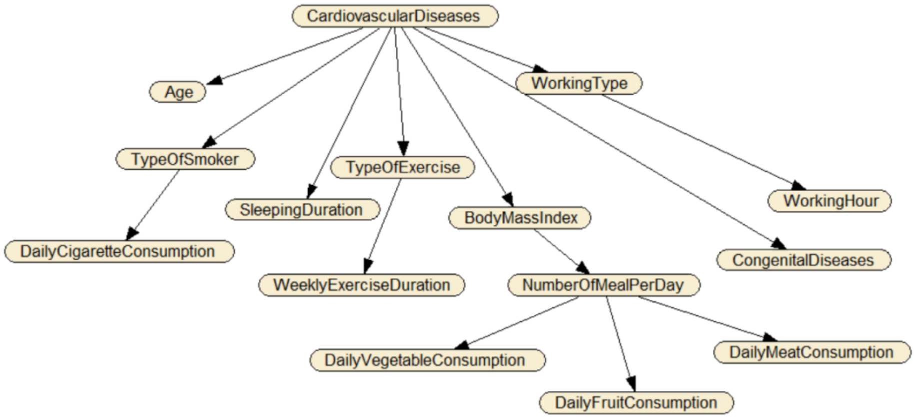

Figure 5. BN Model

3.2. Conditional Probability Table

The aforementioned causative diagram does not naturally possess conditional probability table (CPT). CPT is an embodied probabilistic tables within every node in the BN model. Along with BN model structure, CPT is the calculation basis to gauge the event occurrence in the BN model.

In order to structure CPT, we develop set of questions to attain cardiac profile based on Table 3. Some of questions are listed in Table 4.

Table 4. Some of Questions to Capture Cardiac Profile

No	Questions
1	What kind of heart disease do you have? (choose one) a) Coronary Heart Disease b) Congestive Heart Failure c) Arythmia
2	How old are you?
3	What type of smoker are you? (choose one): a) Active Smoker b) Passive Smoker
4	If you are an active smoker, how many cigarettes do you consume per day? (choose one): a) 1-3 bars b) 4-6 bars c) >6 bars
5	Do you exercise every day? (choose one) a) Yes b) No
6	What kind of exercise do you do? (choose one) a) Aerobic b) Anaerobic
7	How long the duration of exercise in a week? (choose one) a) <6 hours b) 6-10 hours c) >10 hours
8	How many meals do you take per day? a) Less than 3 times b) 3 times c) More than 3 times

We elicit cardiac profile to 30 respondents with cardiovascular disease and we use it as prior probability for CPT. Before we use the result for CPT construction, we test the validity and reliability of the result. For validity test, it can be inferred that the data is valid by comparing the

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

rvalue and rtable. In our case, all rvalue is greater than rtable (0,361). For reliability test, we employ Cronbach Alpha test. The result shows that our data is reliable with the value of test greater than 0.517. The BN Model with CPT inside can be seen in Figure 6. We also run an example of using BN Model to predict the probability of CVD that a person might have. Figure 7 depicts the result of BN Model running. Given someone with number of meal less than three times, active smoker, sleeping duration less than 6 hours per day, and age between 30-60 years old will, she/he have more opportunity to gain coronary heart disease with the probability of 52,2%.

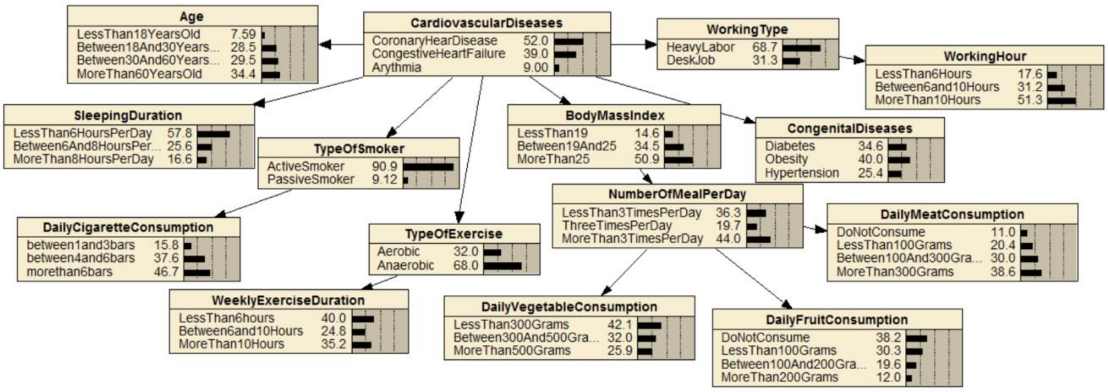

Figure 6. BN Model with CPT Inside

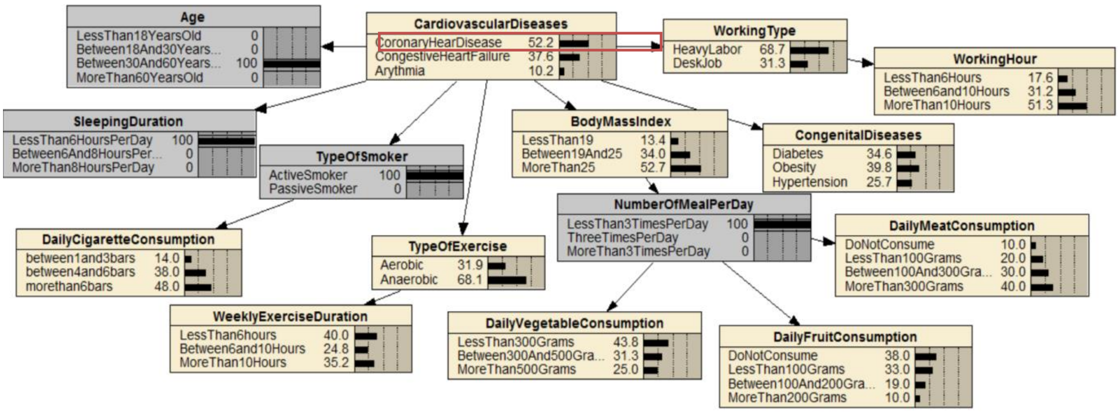

Figure 7. Example of Using BN Model

3.2. Sensitivity Analysis

Bayesian network sensitivity analysis is a method to analyse the degree of coherence among nodes given one node as center. The sensitivity analysis is an extension of traditional entropy analysis in information theory.

By having this sensitivity analysis, we can assess the nodes importance and evaluate which nodes that are more dominant than others. The result of sensitivity analysis is presented in Table 5. Based on sensitivity analysis, we can infer that the most important nodes to contribute to Cardiovascular Diseases node are Congenital Diseases, Type Of Exercise, Body Mass Index, and Age which constitute around five percent.

Table 5. Sensitivity analysis towards “Cardiovascular Diseases” node

Node	Mutual	Percent	Variance of belief
Congenital Diseases	0.01987	01.49	0.0021596
Type O fExercise	0.01767	01.33	0.0047825
Body Mass Index	0.01760	01.32	0.0044392
Age	0.01756	01.32	0.0030089

Sleeping Duration	0.01234	0,643055556	0.0025811
Working Type	0.00678	0,353472222	0.0017895
Type Of Smoker	0.00063	0,327083333	0.0001655
Weekly Exercise Duration	0.00013	0.00969	0.0000360
Working Hour	0.00002	0.00122	0.0000044
Daily Cigarette Consumption	0.00002	0.00113	0.0000041
Number Of Meal Per Day	0.00001	0.000923	0.0000030
Daily Meat Consumption	0.00000	0	0.0000000
Daily Fruit Consumption	0.00000	0	0.0000000
Daily Vegetable Consumption	0.00000	0	0.0000000

4. Conclusion

This paper aims to develop a probabilistic model for predicting the possibility of someone to have cardiovascular disease (CVD) given his/her current lifestyle. We construct our model using Bayesian Network and identify 14 factors causing CVD. We also provide an example on how our model can be used. Furthermore, based on our sensitivity analysis, we find that age, body mass index, congenital diseases, and type of exercise play important role in determining CVD type for each person.

The limitation of this paper is the static nature of the model. In the future, we plan to develop an algorithm to dynamically update the BN model in order to provide better prediction accuracy.

References

[1] "Statistical Fact Sheet 2016 Update," 2016.

[2] N. Townsend, J. Williams, P. Bhatnagar, K. Wickramasinghe, and M. Rayner, *CARDIOVASCULAR DISEASE STATISTICS 2014*. 2014.

[3] "Heart Disease and Stroke Statistics 2017: At-a-Glance," 2017.

[4] "Screening for cardiovascular disease and risk factors," 2011.

[5] A. O. Odegaard, W.-P. Koh, M. D. Gross, J.-M. Yuan, and M. A. Pereira, "Combined Lifestyle Factors and Cardiovascular Disease Mortality in Chinese Men and Women The Singapore Chinese Health Study," *Circulation*, pp. 2847–2854, 2011.

[6] P. Mullie and P. Clarys, "Association between Cardiovascular Disease Risk Factor Knowledge and Lifestyle," *Food Nutr. Sci.*, vol. 2, pp. 1048–1053, 2011.

[7] "Reducing risk in heart disease: an expert guide to clinical practice for secondary prevention of coronary heart disease," 2012.

[8] J. Lv *et al.*, "Adherence to Healthy Lifestyle and Cardiovascular Diseases in the Chinese Population," *J. Am. Coll. Cardiol.*, vol. 69, pp. 1116–1125, 2017.

[9] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early heart disease prediction using data mining techniques," *Comput. Sci. Inf. Technol.*, pp. 53–59, 2014.

[10] J. Vijayashree, N. Ch, and Srimannarayanaiyengar, "Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review," *Int. J. Bio-Science Bio-Technology*, vol. 8, no. 4, pp. 139–148, 2016.