Journal of Quality Measurement and Analysis
Jurnal Pengukuran Kualiti dan Analisis
JQMA 8(2) 2012, 21-36

A CASE STUDY ON QUALITY OF SLEEP AND HEALTH USING BAYESIAN NETWORKS

(Suatu Kajian Kes tentang Kualiti Tidur dan Kesihatan Menggunakan Rangkaian Bayesan)

HONG CHOON ONG, CHIEW SENG LEE & CHYE CHING SIA

ABSTRACT

The objectives of this study are to investigate the associations of the socio-demographic characteristics, living habits, social and interpersonal factors and behaviour-related risk factors with both the quality of sleep and state of health. The tool used is Bayesian networks (BNs), a special case of probabilistic graphical models. The data utilised in this study is collected by employing the sample survey technique through the online network. A total of 1316 sets of data are collected with 20 variables of interest being studied. Our situation is whereby the BNs structure is unknown with full observability. A structural learning is conducted on the data to learn the correct network structure. There are several phases involved including implementation of the learning algorithms, integration of prior knowledge through the whitelist argument and arc setting to form directed acyclic graphs. The network scores computation is implemented to estimate the fitting of the resulting network of each structural learning algorithm in order to choose the best-fitted network. The arc strength or edge intensity is computed to estimate the marginal confidence on the presence of an arc. We found that the quality of sleep is dependent on certain factors in socio-demographic characteristics. The state of health is dependent on some factors in socio-demographic characteristics and living habits. We can also deduce that the quality of sleep has an impact on the state of health.

Keywords: quality of sleep; Bayesian networks; directed acyclic graphs; learning algorithms; network scores

ABSTRAK

Penyelidikan ini bertujuan untuk mengkaji hubungan antara ciri sosiodemografi, gaya hidup, faktor-faktor sosial dan interpersonal, faktor-faktor risiko berkaitan tingkah laku dengan kualiti tidur serta keadaan kesihatan. Kaedah yang digunakan dalam kajian ini adalah rangkaian Bayesan, iaitu salah satu kes khas dalam model grafik kebarangkalian. Data yang digunakan dalam kajian ini dikumpulkan dengan menggunakan teknik tinjauan sampel melalui rangkaian dalam talian. Sejumlah 1316 set data telah terkumpul dengan sebanyak 20 pemboleh ubah yang dikaji dalam projek ini. Situasi penyelidikan ini adalah struktur rangkaian Bayesan tidak diketahui dengan ketercerapan yang lengkap. Pembelajaran struktur rangkaian dijalankan ke atas data untuk mengenal pasti struktur rangkaian yang tepat. Terdapat beberapa fasa yang terlibat, termasuklah pelaksanaan al-Khwarizmi pembelajaran, integrasi pengetahuan yang sedia ada melalui hujah senarai putih dan penetapan lengkok untuk membentuk graf berarah tanpa kitaran. Pengiraan skor rangkaian dilaksanakan untuk menganggar kesesuaian rangkaian yang terhasil bagi setiap algoritma pembelajaran struktur untuk memilih rangkaian yang paling sesuai. Kekuatan atau intensiti lengkok dikira untuk menganggar keyakinan marginal atas kehadiran sesuatu lengkok. Didapati bahawa kualiti tidur adalah bersandaran pada faktor tertentu dalam ciri sosiodemografi. Keadaan kesihatan adalah bersandaran kepada beberapa faktor dalam ciri sosiodemografi dan gaya hidup. Turut disimpulkan bahawa kualiti tidur mempunyai impak ke atas keadaan kesihatan.

Kata kunci: kualiti tidur; rangkaian Bayesan; graf berarah tanpa kitaran; al-Khwarizmi pembelajaran; skor rangkaian

1. Introduction

In recent years, with growing public interest in health issues, increasing concern has been shown over the health-related quality of life. Before moving in depth into the topic and objective of our study which is closely related to health behaviour, we will, first, introduce the concept of health and sleep quality. The traditional biomedical definition of health is simply the absence of disease or physiological malfunction. It is considered incomprehensive as it merely clarified the individual's physiological state in terms of the presence or absence of symptoms of sickness (Weiss & Lonnquist 2003).

Throughout the years, researchers have refined the concept of health in more comprehensive ways. The one which is defined by the World Health Organization is a state of absolute physical, mental and social well-being, unrestricted by the absence of disease or infirmity (World Health Organization 2003). Still, this definition is unable to provide sufficient operational indicators for physical, mental and social health. Hence, the World Health Organization has produced an operational definition of health which is the degree of conformity to the established standards for different demographic groups with normal limits of variation (Health System Performance Assessment 2003).

Now, we moved on to the concept of sleep quality. When one feels adequately safe and secures to reduce control on vigilance and alertness, the optimal susceptible physiological state occurs. This physiological state is known as sleep (Troxel *et al*. 2007). It can also be defined as detachment (perceptually) from and unawareness to the surrounding in an active, rhythmic and reversible state (Dewald *et al.* 2010). Sleep is considered as a product of central nervous system that generally facilitates normal neuronal function and may contribute positively to general health (Lee 2006).

In terms of sleep quality, it is the experience of sleep including the rested feeling after waking up and sleep satisfaction (Dewald *et al.* 2010). Sometimes, it is characterised by sleep measures such as the total sleep time (TST), sleep onset latency (SOL) (time required from full wakefulness to the onset of sleep), total wake time and sleep efficiency (Krystal & Edinger 2008).

The tool used in this study is the Bayesian network (BN). BNs combine the principles of graph theory, probability theory, statistics and computer science. The BN structures formed are a representation of the knowledge of uncertain domain based on the probability.

BNs are defined by B = (G, Θ) with G representing the directed acyclic graph (DAG) with the graphical structure G = (**V**, A), where **V** is the node or vertex (random variable) set and A is the arc or edge (dependence among the variables) set. The graph G denotes the independence assumptions whereby each variable (node a_i) is independent of its non-descendents given its parents. Θ is defined as the set of parameters of the network. The parameter implied here is $\theta = P(a_i | P(a_i' s\ Parents)$ (Ben-Gal 2007).

BNs take the form of a DAG that represents a joint probability distribution. The directed and acyclic features are crucial to the factorisation of joint probability of a collection of random variables. An edge which traversed from node a_i (parent) to node a_j (child) indicates a statistical dependence among the corresponding variables. This indicates that BNs can be used to represent any joint probability distribution (Pourret 2008). The intrinsic relationship between the variables is formulated in Eq. which is developed from the joint probability of n attributes a_i (chain rule in probability theory) (Witten & Frank 2005).

$$P(a_1, a_2, ..., a_n) = \Pi_{i=1}^{n} P(a_i | a_{i-1}, ..., a_1) = \Pi_{i=1}^{n} P(a_i | a_i' s\ Parents) \quad (1)$$

Basically, there are few types of BN learning cases which are often being considered (Ben-Gal 2007). However, only the situation whereby the BN structure is unknown without missing data is considered in this study.

The main objective of our study is to investigate the associations of the four components which are the socio-demographic characteristics, living habits, social and interpersonal factors and behaviour-related risk factors with both the quality of sleep and state of health.

We are also interested in exploring the relationship among the socio-demographics characteristics, living habits, social and interpersonal factors and behaviour-related risk factors. Moreover, the interaction between the quality of sleep and state of health is being studied. We used BNs to establish the causal relationship (dependencies) among the four components in our study and the two outcome variables and identify the intensity or strength of the established relationships.

2. Literature Review

BNs are widely used in various fields due to its ability to provide reasoning based on real world model. BNs allow a complete understanding on the process involved. A combination of this criteria and strong probabilistic theory enable BNs to produce an objective interpretation. Owing to the nature of BNs, causal relationships among the variables can be learned and hence enhancing the prediction performance. BNs have both causal and probabilistic semantics which makes it a powerful representation of the data. Larose (2006) stated that BNs allow joint conditional independencies to be defined among subsets of variables. BNs have been applied in the studies in automotive industry by Cinicioglu *et al.* (2012), organisational behaviours and management by Oliveira *et al.* (2012) and socio economic by Ge *et al.* (2010).

Various studies on health behaviours and sleep performance have been performed to discover the associations among the health-related attributes. Strine and Chapman (2005) have performed a study on the associations of frequent sleep insufficiency with health-related quality of life (HRQOL). They used the bivariate analyses to investigate the unadjusted prevalence of socio-demographic characteristics, HRQOL, and adverse health behaviours by sleep status and the logistic regression analyses to identify potential covariates of frequent sleep insufficiency. The results indicated that sleep insufficiency is the major health problem. Dalstra *et al*. (2006) have carried out a study on a comparative appraisal of the relationship of socio-economic variables with less than good health among the elderly. The results showed that there are impacts of socio-economic on health differences among the elderly. Barclay *et al*. (2011) applied the structural equation modelling to show the extent to which the genetic and environmental effects explained the association between dependent negative life events and sleep quality. The results showed that the association between the dependent negative life events and sleep quality is great and there are substantial genetic influences on dependent negative life events.

Although there are many studies that have been established on health and sleep performance issues which focused on investigation the association among the variables, they mainly apply the traditional statistical tools such as regression analysis. There are limitations to these methods. These methods are not suitable to be applied in typical nonlinear behaviour and where the normality condition has to be fulfilled. However, these constraints can be overcome by the use of BNs. BNs are particularly useful in establishing relationships that are highly variable. Moreover, BNs are characterised by the nonlinearity of the analyzed system (Oliveira *et al*. 2012). Thus, it can be utilised to explore the relationships among the variables on the area of health and sleep research.

3. Data and Methodology

3.1. *Research Variables*

The state of health and the quality of sleep are two crucial components in health-related quality of life. There are four components of factors and two final outcomes being considered in this study. The four groups of factors are socio-demographic characteristics, living habits, social and interpersonal factors and behaviour-related risk factors whereas the two final outcomes are the state of health and the quality of sleep.

The measurement approach of state of health adopted in this study is the self-rated overall health status. It is a common health indicator with proven validity despite being subjective. It also possessed good test-retest reliability (Dalstra *et al*. 2006). The health status defined here considered the operational measures, thus, it can be referred as a degree of conformity to the established standards for different demographic or social groups with normal limits of variation.

The other outcome of this study is the quality of sleep which is referred as the individual evaluation of sleep performance including the sleep satisfaction as well as the feeling of restfulness after waking up. This definition is a modified version of the one given by Dewald *et al.* (2010). Previous studies have shown that sleep performance is often associated with a person's ability to function optimally and psychological well-being (Strine & Chapman 2005).

The first group of factors is characterised by socio-demographic which is a combination of sociological and demographic characteristics. These factors are gender, age, body weight and height, Body Mass Index (BMI), blood type, family history of diseases, marital status, level of education and living environment. The BMI is a commonly used predictor of the healthy levels of weight for height among individuals (Henderson 2005). The level of education is included in this study as it is associated with the proficiency to access to information and benefit from these new knowledge (Dalstra *et al*. 2006). It is thus believed to have an influence individual health and sleep performance.

The second component is the living habits. In this study, we defined living habits as the routine habits that are significant in health. The factors that are grouped in this component are daily water consumption, preserved food consumption and regularity of performing physical activities or exercises.

The third component is the social and interpersonal factors. These are related to the emotional behaviours and responses of an individual. The factors are frequencies of emotional problem faced by an individual, frequencies of emotional support received by an individual and rate of participation in recreational activities. According to Troxel *et al.* (2007), sleep and social environment are often referred as the ability to control emotions and behavioural responses to interpersonal affairs.

The fourth group of factors is related to the behavioural-related risk factors. The behavioural-related risk factors adopted in this study is defined as the attributes or characteristics that increase the likelihood of developing disease (Health System Performance Assessment 2003). These factors are the habit of smoke and consumption of alcohol.

3.2. *Research Data*

The sample survey technique is applied in the data collection whereby a survey questionnaire is conducted. The Health-Related Quality of Life Survey Questionnaires (2011) is established. The questionnaires are set by referring to the Behavioural Risk Factor Surveillance System (BRFSS) Questionnaire (2011) and Health System Performance (HSP) Assessment (2003). The

BRFSS Questionnaire is a surveillance system conducted by the health department with the support from Centers for Disease Control and Prevention (CDC) in United States. The HSP Assessment is a World Health Survey programme conducted in India and initiated by the World Health Organization (WHO). The questionnaires consist of 19 questions which are divided into five sections.

The questions in this survey questionnaire require the respondents to evaluate their health-related performance based on individual perception. The target population of this study is the internet users from Malaysia with age ranging from 10 to 50 years old. Hence, the questionnaires are sent to the respondents via social network and email (electronically) to obtain the sample population. The sample design applied in this study is the convenience sampling method because it is practical as it is inexpensive and subjects are readily available and it also provides evidence of the occurrence of particular quality of an event within a given sample. The raw data collected are utilised as the data of our study. The data collected is being processed before it is analyzed. After eliminating the duplicated data, we have obtained a total of 1,316 sets of data.

There are 16 variables (state of health, quality of sleep, gender, level of education, living environment, blood type, family history of diseases, marital status, daily water consumption, preserved food consumption, regularity of performing physical activities or exercises, frequencies of emotional problems, frequencies of emotional support, rate of participation in recreational activities, habit of smoke and the consumption of alcohol).

A new variable, the Body Mass Index (BMI) is formed by utilizing the data for weight and height using the standard formula for calculating BMI (Henderson 2005) as shown in Eq. .

$$BMI = \frac{\text{Weight in kilograms}}{\text{Height in meters}^2} \tag{2}$$

The completed data set will be utilised as an input data for structural learning in Bayesian Networks (see Table 1).

Table 1: The Variables and the Levels of Response for Each Variable

Component and factors	Variable	In Short	Level of Response	Response Description
Outcomes	State of Health	Health	5	Very Bad, Bad, Moderate, Good, Very Good
	Quality of Sleep	Sleep	5	None, Mild, Moderate, Severe, Extreme
Socio- Demographic Characteristics	Gender	Gender	2	Male, Female
	Age	Age	4	A, B, C, D
	Weight	Weight	5	A, B, C, D, E
	Height	Height	6	A, B, C, D, E, F
	Body Mass Index	BMI	3	Underweight, Normal Weight, Overweight
	Level of Education	Edu.	5	Primary School, Secondary School, High School or Equivalent, College or University, Post Graduate Degree
	Living Environment	Env.	3	Rural, Semi-Urban, Urban
	Blood Type	Blood	5	A, B, AB, O, Don't Know
	Family History of Diseases	DisH	3	Yes, No, Don't Know
	Marital Status	Marital	2	Single, Currently Married
Living Habits	Daily Water Consumption	Water	4	"$x < 1$", "$1 \leq x < 2$", "$2 \leq x < 3$", "$x \geq 3$"
	Preserved Food Consumption	Food	5	Never, Sometimes, About Half the Time, Most of the Time, Always
	Regularity of Performing Physical Activities or Exercises	Exe.	5	Never, Once In A While, Monthly, Weekly, Daily
Interpersonal and Social	Frequencies of Emotional Problem	EmoP	5	None, Mild, Moderate, Severe, Extreme
	Frequencies of Emotional Support	EmoS	5	Never, Rarely, Sometimes, Usually, Always
	Rate of Participation in Recreational Activities	Rec.	5	Never, Rarely, Sometimes, Usually, Always
Behaviour-Related Risk Factors	Habit of Smoke	Smoke	3	No; Yes, But Not Daily; Yes, Daily
	Consumption of Alcohol	Alcohol	5	Never, Once In A While, Monthly, Weekly, Daily

3.3. *Research Methods*

The structure learning is a method that encodes the conditional independence present in the data. A distribution that coincide or is as close as possible to the actual distribution in the probability space will be identified by utilizing this method (Scutari 2011b). Basically, to learn the network structure, several phases are implemented which include the search through a space consisting of all possible sets of edges, estimation of the conditional probability tables for each set and computation of the log-likelihood of the resulting network based on the data as a measure of the network's quality (Witten & Frank 2005). The computation of BNs is conducted by using the R Language version 2.14.0. Both bnlearn and Rgraphviz packages are used; the former is used for the structural learning of BNs whereas the latter offers advance plotting options of the resultant networks (Scutari 2010).

The structural learning algorithms are used to construct BNs for the data. These algorithms are developed to learn the correct network with reduced computational complexity. Computational complexity occurs when the data sets have high dimensionality (Scutari 2010). There are, basically, three approaches of algorithms which will be applied in our study; the constraint-based algorithms, score-based algorithms and hybrid structural learning algorithms. The Grow-Shrink (GS), Incremental Association Markov Blanket (IAMB), Fast Incremental Association (Fast-IAMB) and Interleaved Incremental Association (Inter-IAMB) algorithms are types of constraint-based algorithms. The Hill-Climbing (HC) and Tabu Search (Tabu) algorithm are types of score-based algorithms. The Max-Min Hill-Climbing (MMHC) and General 2-Phase Restricted Maximisation (RSMAX2) algorithm are types of hybrid structural learning algorithms (Scutari 2011a).

These algorithms differ in the method they search through the network structures space. The constraint-based algorithms learn the network structure by analyzing the probabilistic relations with conditional independence tests. These relations are entailed by the Markov property (an application of chain rule of probability) of BNs. Then, a graph will be constructed. The resulting models are usually identified as causal models. The score-based algorithms are general optimisation techniques that function by allocating a score indicating goodness of fit to each candidate of BNs and these score are maximised with some heuristic search algorithms (algorithms that explore the search space by starting with the network structure and adding, deleting and reversing one arc at a time until the score cannot be improved). The hybrid structural learning algorithms combine the constraint-based and score-based algorithms to counterbalance the limitations of both the algorithms (Scutari 2011b).

The scoring functions are implemented on the directed graph of each learning algorithm to measure the quality of a given network. These network scores are computed to estimate the fitting of structural learning algorithms and thus, the algorithms which fit the data best is chosen. These scoring functions operate by balancing the fit to the data with the complexity of the model in order to avoid over fitting (Ge *et al*. 2010).

The complexity occurs because the quality is measured by calculating the probability of the data given the network (the probability of network according to each instance is calculated and these probabilities are multiplied over all instances) which yielded numbers that are too small to represent the data well. Instead, the sum of the logarithms of the probabilities is used. The value yielded is the log-likelihood of the network given data. If the log-likelihood is maximised based on the data, the network formed will be over fitted. Thus, a penalty is added for the complexity of the network based on the number of parameters (total number of independent estimates in all the probability tables) to prevent over fitting (Witten & Frank, 2005). It is expressed in the form of Eq. (3) (Ge *et al*. 2010).

$$Score(G,D) = \log \hat{P}(D \mid G) - \Delta(D,G) \tag{3}$$

where G represents the directed acyclic graph, D represents certain data set and $\Delta(D,G)$ represents the complexity penalty.

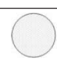
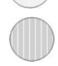
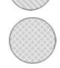
There are a few network scores that are implemented which are suitable for discrete data usage. These score functions include the log-likelihood (loglik), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Bayesian Dirichlet Equivalent (BDE), and Cooper & Herskovits' K2 score (Scutari 2011b).

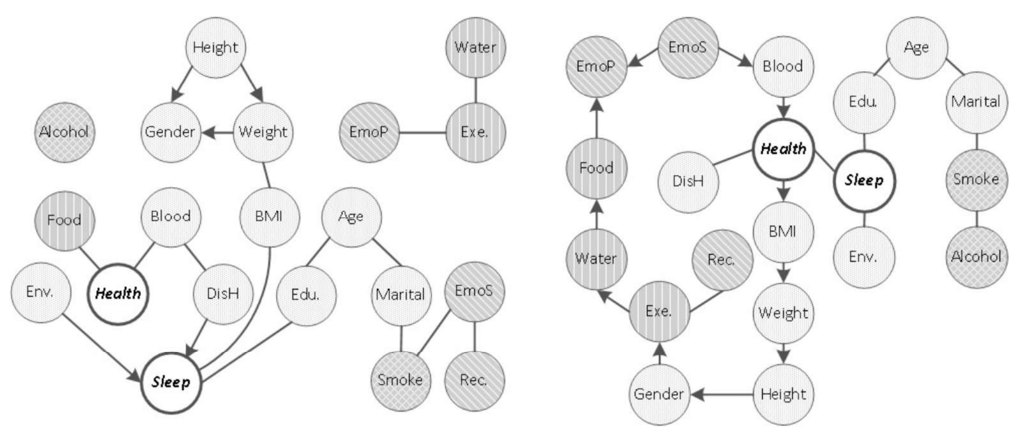
4. Analysis, Result and Discussions

4.1. *Analysis*

By implementing the learning algorithms available in the bnlearn package, eight different networks are formed (see Figure 1 and Table 2). The existence of arcs or edges indicates the presence of dependent relationship between the variables (nodes) whereas the absence of arcs or edges indicates the presence of conditional independence between the variables. From these networks, it can be seen that there exist a strong relationship between some of the variables as the arcs or edges between these particular variables are present in almost all the eight networks.

Table 2: The Indication of the Pattern of Each Node.

Colour	Component
	Socio-demographic Characteristics
	Living Habits
	Social and Interpersonal
	Behaviour-Related Risk Factors

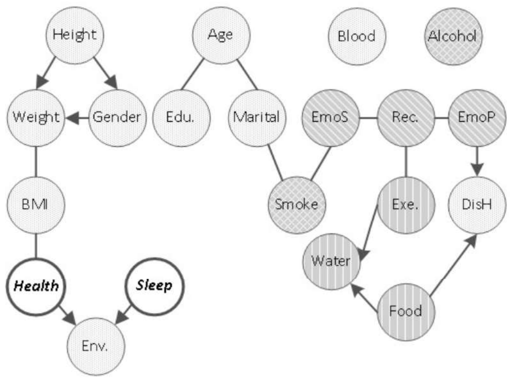


(a) Grow-Shrink (GS) Learning Algorithm.

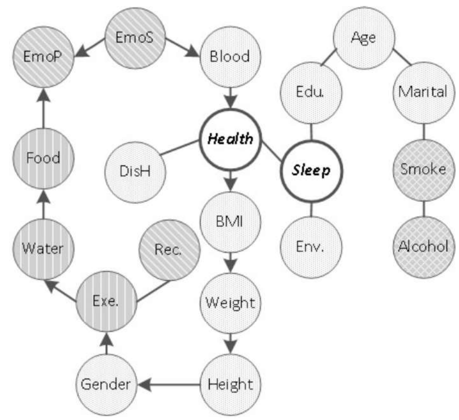
(b) Incremental Association Markov Blanket (IAMB) Learning Algorithm.

To be continued...

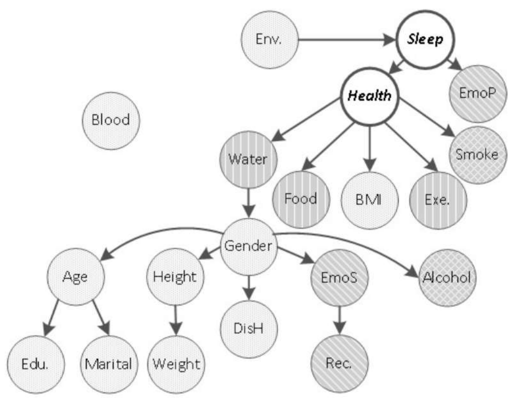
...Continuation



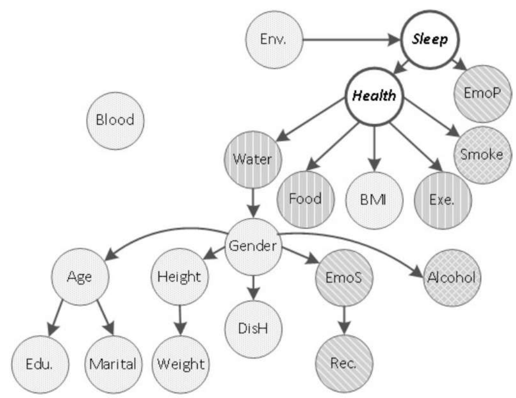
(c) Fast Incremental Association (Fast-IAMB) Learning Algorithm



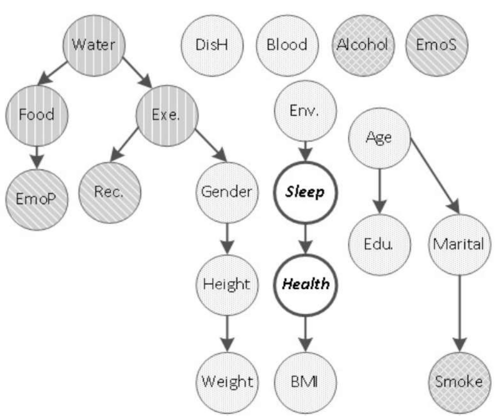
(d) Interleaved Incremental Association (Inter-IAMB) Learning Algorithm



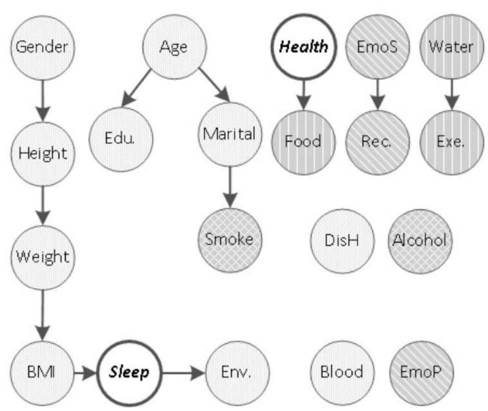
(e) Hill-Climbing (HC) Learning Algorithm



(f) Tabu Search (Tabu) Learning Algorithm



(g) Max-Min Hill-Climbing (MMHC) Learning Algorithm



(h) General 2-Phase Restricted Maximisation (RSMAX2) Learning Algorithm

Figure 1: Initial Networks Using Various Learning Algorithms

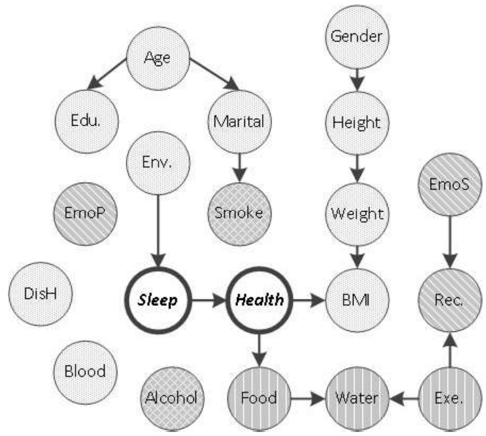
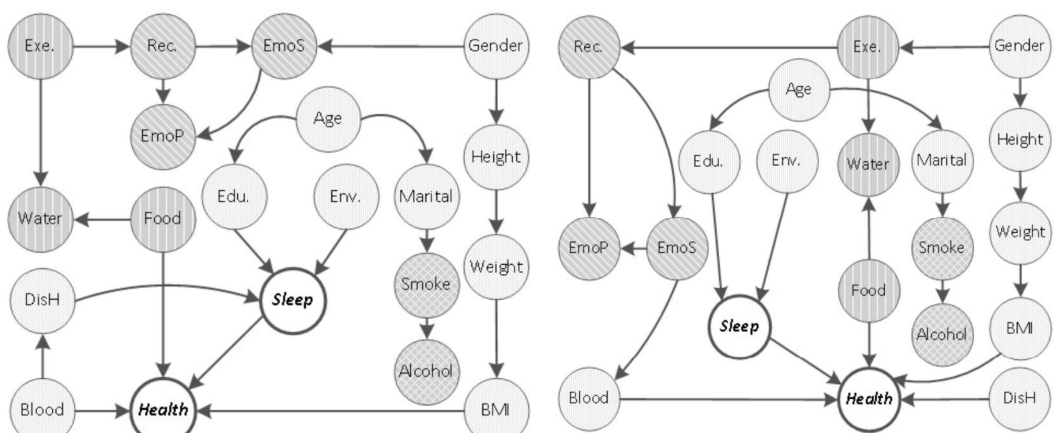


Figure 2: The Network for Whitelisting from the Common Arcs in the 8 Networks

The arcs which exist in at least half of the eight networks and present in at least two out of the three types of algorithms (i.e. constraint-based, score-based and hybrid structural learning algorithms) indicating strong associations among the variables. Thus, these arcs are set as whitelist (see Figure 2). Once again, the learning algorithms are run using the white listed arcs. After the changes and directions are determined, these white listed algorithms are run again with the addition of arc setting. Consequently, all eight of the networks have transformed into directed acyclic graphs (see Figure 3).

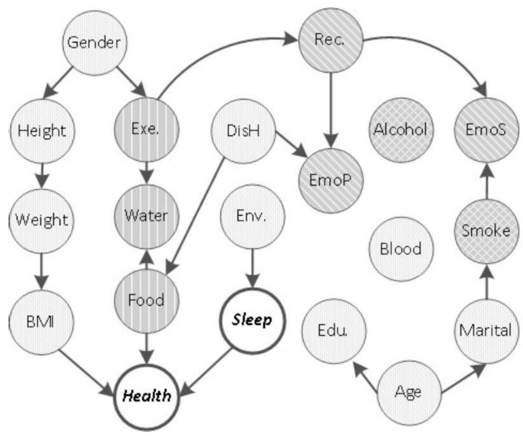


(a) Grow-Shrink (GS) Learning Algorithm

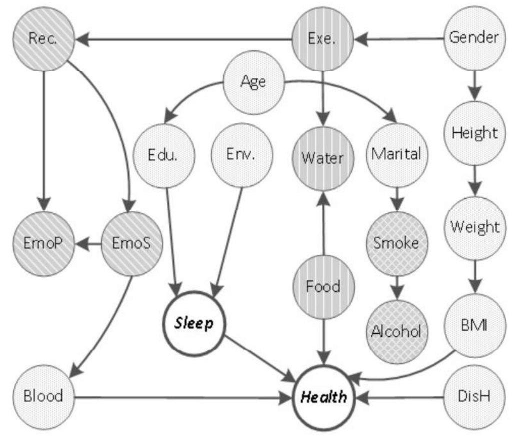
(b) Incremental Association Markov Blanket (IAMB) Learning Algorithm

To be continued...

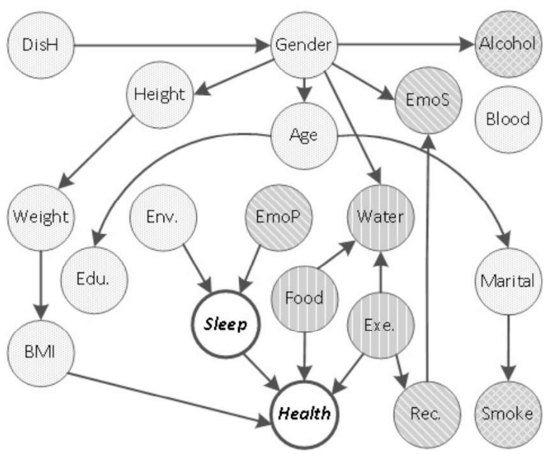
...Continuation



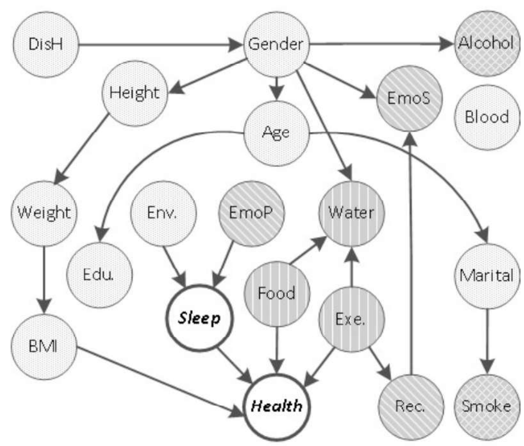
(c) Fast Incremental Association (Fast-IAMB) Learning Algorithm



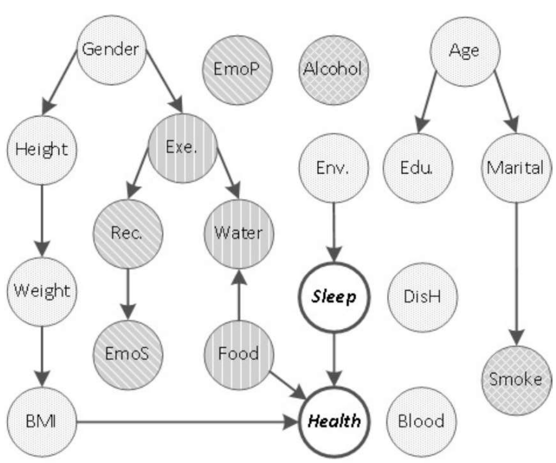
(d) Interleaved Incremental Association (Inter-IAMB) Learning Algorithm



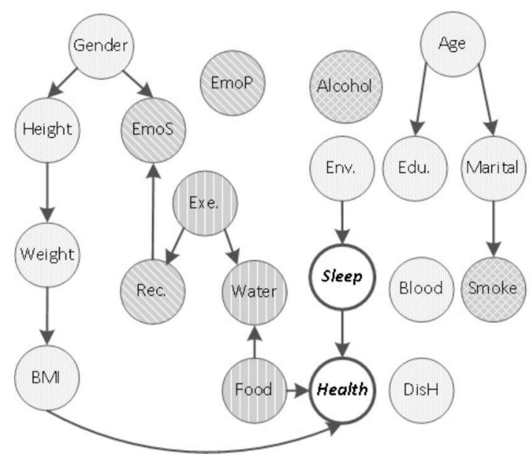
(e) Hill-Climbing (HC) Learning Algorithm



(f) Tabu Search (Tabu) Learning Algorithm



(g) Max-Min Hill-Climbing (MMHC) Learning Algorithm



(h) General 2-Phase Restricted Maximisation (RSMAX2) Learning Algorithm

Figure 3: The Final Directed Networks from the Various Learning Algorithms

Upon obtaining the directed acyclic network, we proceed to the computation of the network score of the eight directed networks formed (see Table 3). The network obtained by MMHC algorithm has the highest score in three out of five network scores i.e. BDE score, Cooper & Herskovits' K2 score and BIC score. Thus, we can conclude that this network performed well in almost all of the network scores and it is selected as our final network.

Table 3: The Network Score of Each Learning Algorithm.

Learning Algorithms \ Network Score	bde	k2	loglik	aic	bic
Grow-Shrink (GS)	-26554.14	-26825.78	-24160.47	-27008.47	-34388.14
Incremental Association Markov Blanket (IAMB)	-27124.41	-27619.48	**-24072.33**	-31293.33	-50004.22
Fast Incremental Association (Fast-IAMB)	-25756.82	-25846.84	-24480.06	**-25327.06**	-27521.79
Interleaved Incremental Association (Inter-IAMB)	-27124.41	-27619.48	**-24072.33**	-31293.33	-50004.22
Hill-Climbing (HC)	-26232.31	-26756.25	-24089.47	-27272.47	-35520.18
Tabu Search (Tabu)	-26232.31	-26756.25	-24089.47	-27272.47	-35520.18
Max-Min Hill-Climbing (MMHC)	**-25637.54**	**-25800.63**	-24601.6	-25330.6	**-27219.57**
General 2-Phase Restricted Maximisation (RSMAX2)	-25671.74	-25815.6	-24597.9	-25341.9	-27269.73

[a]the bold figures indicate the highest score for each score function.

Next, we proceed to the computation of the arc strength of this network (see Table 4). The higher the value indicated that the higher the strength of association. The table showed that the intensity of the strength from the top towards the bottom in a decreasing manner. The intensity of the relationship is very strong between Food and Health, BMI and Health, Sleep and Health, Food and Water as well as Exe. and Water. The intensity of the strength of arc is transformed into graphical formed as shown in Figure 4. Notice that the thicker the arc, the stronger the association among the variables.

Table 4: The Network Score of Each Learning Algorithm.

From	To	Strength
Food	Health	1022.4042
BMI	Health	997.4735
Sleep	Health	976.7828
Food	Water	268.1682
Exe.	Water	263.6774
Exe.	Rec.	-1.1536
Rec.	EmoS	-1.5846
Marital	Smoke	-1.6826
Env.	Sleep	-8.1753
Age	Edu.	-9.4112
Gender	Exe.	-15.4037
Weight	BMI	-50.4315
Age	Marital	-69.1001
Height	Weight	-198.5008
Gender	Height	-364.3684

4.2. *Result and Discussions*

Since the sampled population of our study is internet users from Malaysia with age ranging from 10 to 50 years old, our result portrayed the associations between the variables based on the Malaysia's internet users. This implied that the focus of our discussion is specifically on mainly internet user population from Malaysia.

From Figure 4 and Table 4, it can be seen that the directions of strong arcs are from food, BMI and sleep to health with the strengths of dependency 1022.4042, 997.4735 and 976.7828 respectively. These imply that the strengths of dependency are very strong. Hence, we can conclude that health is strongly and directly dependent on food, BMI and sleep. This indicated that the state of health of Malaysia's internet users is influenced by preserved food consumption, quality of sleep and BMI. The association between quality of sleep and state of health is corroborated by the study by Strine and Chapman (2005) which established that insufficient sleep is the major problem of health.

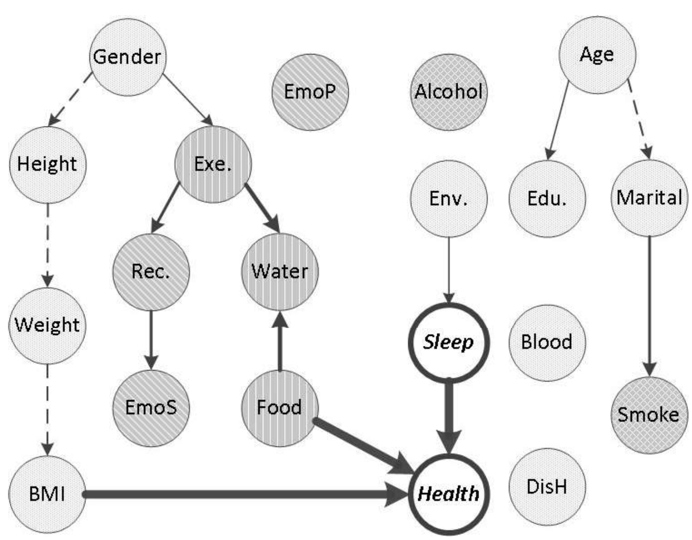


Figure 4: The Final Network with the Various Arc Strengths

The dependency of state of health on preserved food consumption showed that there exists relationship between these variables. This is supported by the Health System Performance Assessment (2003) which mentioned that the nutritional content of food is directly associated with health and the use of tobacco and alcohol have detrimental effects on health. We can conclude that living habits are associated with the state of health.

However, in our study, the alcohol consumption does not portray any association with other variables and the status of smoke does not demonstrate any dependency both directly and indirectly on state of health. These are due to the non-response error in which the majority of the respondents are non-alcoholics (93.47% of our respondents never or consume alcohol once in a while) and non-smokers (89.82% of our respondents never smoke). Thus, the sample does not represent the population well in these particular attributes.

Aside from the direct relationship between BMI and state of health, there are indirect and less intense relationships that exist whereby the state of health is indirectly dependent on gender, height, weight and living environment. These relationships indicated that there is certain association between socio-demographic characteristics and state of health. Also, the fact that there exist association between BMI and state of health is corroborated by Henderson (2005) which presented that BMI is an accurate indicator of health.

The quality of sleep is directly dependent on living environment. This fact is confirmed by the studies conducted by Strine and Chapman (2005) which verified that health-related quality of life factors are associated with the quality of sleep. Thus, we can conclude that the quality of sleep is dependent on the socio-demographic characteristics.

Moreover, daily water consumption is dependent directly and moderately on both preserved food consumption and regularity of performing physical activities or exercises and indirectly and weakly on gender. Regularity of performing physical activities or exercises is depended on by rate of participation in recreational activities (directly) and by frequencies of emotional support received by an individual (indirectly). These dependencies specified that there are certain associations between the socio-demographic characteristics, living habits and social and interpersonal factors whereby the social and interpersonal factors is directly dependent on living habits and indirectly on socio-demographic characteristics.

The habit of smoking is dependent on marital status (directly) and age (indirectly) indicating that there is certain association between socio-demographic characteristics and behaviour-related risk factors. There seems to be no relationship being established for frequencies of emotional problem, blood type and family history of diseases due to the absence of arcs suggesting that these variables are not significant in affecting the quality of sleep and state of health for internet users.

5. Conclusion

The Bayesian networks (BNs) have gained its popularity due to its ease of computation and understanding. Besides, it is suitable for large and highly variable data set which is why it is suitable for our study. In this study, we can conclude that the result of our study showed that there are evident associations between the socio-demographic characteristics, living habits and social and interpersonal factors. We also found that there are certain associations between socio-demographic characteristics and behaviour-related risk factors. Besides, there exists certain association between the quality of sleep with the socio-demographic characteristics whereby the quality of sleep is dependent on these factors. Furthermore, we also found that there are associations between the state of health with the socio-demographic characteristics and living habits. The state of health is dependent on these factors both directly and indirectly. Lastly, we can deduce that there is an association between state of health and quality of sleep whereby the quality of sleep have an impact on state of health. These results were in line with the studies on health-related quality of life conducted previously.

Acknowledgments

This work is partially supported by Universiti Sains Malaysia short term grant number 304/ PMATHS/6311126.

References

Barclay N. L., Eley T. C., Rijsdijk F. V. & Gregory A. M. 2011. Dependent negative life events and sleep quality: an examination of gene-environment interplay. *Sleep Medicine* **12**: 403-409.

Behavioral Risk Factor Surveillance System (BRFSS) Questionnaire. 2011. http://www.cdc.gov/brfss/questionnaires/pdf-ques/2011brfss.pdf (10 December 2011).

Ben-Gal I. 2007. Bayesian networks. *Encyclopedia of Statistics in Quality and Reliability,* New Jersey: John Wiley & Sons.

Cinicioglu E., Onsel S. & Ülengin F. 2012. Competitiveness analysis of automotive industry in Turkey using Bayesian networks. *Expert System with Applications* **39**: 10923-10932.

Dalstra J. A. A., Kunst A. E. & Mackenbach J. P. 2006. A comparative appraisal of the relationship of education, income and housing tenure with less than good health among the elderly in Europe. *Social Science and Medicine* **62**: 2046-2060.

Dewald J. F., Meijer A. M., Oort F. J., Kerkhof G. A. & Bögels S. M. 2010. The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescent: a meta – analytic review. *Sleep Medicine Reviews* **14**: 179-189.

Ge Y., Li C. & Yin Q. 2010. Study on factors of floating women's income in Jiangsu province based on Bayesian networks. *Adv. in Neural Network Research & Application* **67**: 819-827.

Health-Related Quality of Life Survey. 2011. https://docs.google.com/spreadsheet/embeddedform?formkey=dFhGRHU5NUhYeU9UMGQ0MHNCOWxfT0E6MQ (11 May 2012).

Health System Performance Assessment. 2003. http://www.who.int/healthinfo/survey/whs_hspa_book.pdf (10 December 2011).

Henderson R. M. 2005. The bigger the healthier: are the limits of BMI risk changing overtime? *Economics and Human Biology* **3**: 339-366.

Krystal A. D. & Edinger J. D. 2008. Measuring sleep quality. *Sleep Medicine* **9 Suppl. 1**: S10-S17.

Larose D. T. 2006. *Data Mining Methods and Models*. New Jersey: John Wiley & Sons Inc.

Lee T. 2006. *Sleep: A Comprehensive Handbook*. New Jersey: John Wiley & Sons Inc.

Oliveira M. A., Possamai O., Valentina L. V. O. D. & Flesch C. A. 2012. Applying Bayesian networks to performance forecast of innovation projects: a case study of transformational leadership influence in organizations oriented by projects. *Expert System with Applications* **39**: 5061-5070.

Pourret O. 2008. Introduction to Bayesian networks. In Pourret O., Naim P. & Marcot B. (eds.). *Bayesian Networks: A Practical Guide to Applications:* 1-13. West Sussex: John Wiley & Sons Ltd.

Scutari M. 2010. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software* **35** (3).

Scutari M. 2011a. bnlearn-an R package for Bayesian network learning and inference. http://www.bnlearn.com/ (10 December 2011).

Scutari M. 2011b. Measures of variability for graphical models. PhD Thesis. University of Padova.

Strine T. W. & Chapman D. P. 2005. Associations of frequent sleep insufficiency with health-related quality of life and health behaviors. *Sleep Medicine* **6**: 23-27.

Troxel W. M., Robles T. F., Hall M. & Buysse D. J. 2007. Marital quality and the marital bed: examining the covariation between relationship quality and sleep. *Sleep Medicine Reviews* **11**: 389-404.

Weiss G. L. & Lonnquist L. E. 2003. *The Sociology of Health, Healing, and Illness*. 4th Ed. New Jersey: Pearson Education Inc.

Witten I. H. & Frank E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Ed. San Francisco: Morgan Kaufmann Publishers.

World Health Organization. 2003. WHO definition of health.
http://www.who.int/about/definition/en/print.html (30 April 2012).

School of Mathematical Sciences
Universiti Sains Malaysia
11800 USM Penang
MALAYSIA
E-mail: hcong@cs.usm.my, lcs105904@student.usm.my, c.ching_89@hotmail.com*

* *Corresponding author*