Mathematical Biosciences and Engineering

MBE, 19(12): 13660–13674.
DOI: 10.3934/mbe.2022637
Received: 11 July 2022
Revised: 28 August 2022
Accepted: 30 August 2022
Published: 19 September 2022

http://www.aimspress.com/journal/MBE

Research article

Using Bayesian network model with MMHC algorithm to detect risk factors for stroke

Wenzhu Song[1], Lixia Qiu[1], Jianbo Qing[2], Wenqiang Zhi[2], Zhijian Zha[3], Xueli Hu[1], Zhiqi Qin[4], Hao Gong[4] and Yafeng Li[2,5,6,7,*]

[1] School of Public Health, Shanxi Medical University, Taiyuan, China
[2] Department of Nephrology, Shanxi Provincial People's Hospital (Fifth Hospital) of Shanxi Medical University, Taiyuan, China
[3] Chinese Internal Medicine, Shanxi University of Chinese Medicine, Taiyuan, China
[4] Department of Biochemistry & Molecular Biology, Shanxi Medical University, Taiyuan, China
[5] Core Laboratory, Shanxi Provincial People's Hospital (Fifth Hospital) of Shanxi Medical University, Taiyuan, China
[6] Shanxi Provincial Key Laboratory of Kidney Disease, Taiyuan, China
[7] Academy of Microbial Ecology, Shanxi Medical University, Taiyuan, China

* **Correspondence:** Email: dr.yafengli@gmail.com; Tel: +13935151151.

Abstract: Stroke is a major chronic non-communicable disease with high incidence, high mortality, and high recurrence. To comprehensively digest its risk factors and take some relevant measures to lower its prevalence is of great significance. This study aimed to employ Bayesian Network (BN) model with Max-Min Hill-Climbing (MMHC) algorithm to explore the risk factors for stroke. From April 2019 to November 2019, Shanxi Provincial People's Hospital conducted opportunistic screening for stroke in ten rural areas in Shanxi Province. First, we employed propensity score matching (PSM) for class balancing for stroke. Afterwards, we used Chi-square testing and Logistic regression model to conduct a preliminary analysis of risk factors for stroke. Statistically significant variables were incorporated into BN model construction. BN structure learning was achieved using MMHC algorithm, and its parameter learning was achieved with Maximum Likelihood Estimation. After PSM, 748 non-stroke cases and 748 stroke cases were included in this study. BN was built with 10 nodes and 12 directed edges. The results suggested that age, fasting plasma glucose, systolic blood pressure, and family history of stroke constitute direct risk factors for stroke, whereas sex, educational levels, high density lipoprotein cholesterol, diastolic blood pressure, and urinary albumin-to-creatinine ratio

represent indirect risk factors for stroke. BN model with MMHC algorithm not only allows for a complicated network relationship between risk factors and stroke, but also could achieve stroke risk prediction through Bayesian reasoning, outshining traditional Logistic regression model. This study suggests that BN model boasts great prospects in risk factor detection for stroke.

Keywords: stroke; Bayesian network; logistic regression; risk factors; model construction

Abbreviations: BN: Bayesian network; PSM: Propensity core matching; DAG: Directed acyclic graph; CPT: Conditional probability tables; MMHC: Max-Min Hill-Climbing; CKD: Chronic kidney disease; BMI: Body mass index; HDL-C: High-density lipoprotein cholesterol; LDL-C: Low-density lipoprotein cholesterol; TC: Total cholesterol; TG: Triglyceride; GHb: Glycosylated haemoglobin; FPG: Fasting plasma glucose; IFG: Impaired Fasting Glucose; Hcy: Homocysteine; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; MALB: Urinary microalbumin; α_1MG: Urinary 1-microglobulin; ACR: Urinary albumin-to-creatinine ratio; MCR: Urinary 1-microglobulin-to-creatinine ratio

1. Introduction

Stroke is a major chronic non-communicable disease that seriously harms one's health, with high incidence, high mortality, and high recurrence [1]. It's the second leading cause of death and a major cause of disability worldwide [2]. Despite substantial reductions in age-standardised rates, the annual number of strokes and deaths due to stroke increased substantially from 1990 to 2019, posing an alarming burden for the national healthcare system. Also, its prevalence is increasing rapidly in China with over 2 million new cases annually [3], more in rural (298 cases per 100,000 person-years) than urban areas (204 cases per 100,000 person-years) [3,4], more in northern than southern areas [5] and its burden ranks first globally [6], especially in her rural areas that are subjected to ageing populations and underused medical facilities. Yet, it's hard to arouse public attention for its unobvious early symptoms and sudden onset. Obviously, it has emerged as a public health issue. To comprehensively digest its risk factors and thus take some relevant measures to reduce its prevalence are all the more important.

Previously, Logistic regression was employed to explore the risk factors of stroke [7,8], suggesting that hypertension, hyperlipidemia, diabetes mellitus, coronary heart disease and smoking are significantly associated with stroke. Some drawbacks, however, come with the model. The first one concerns independent variables [9]. In clinical research, correlation often exists in risk factors, but the model is unable to meet the prerequisite of independence between variables. The second one lies in its inability to make a sequential prediction. Unavailability of data is common in clinical research, and this interferes with model functionality. The third one is that the model fails to identify direct or indirect risk factors. A better model is needed.

Bayesian networks (BNs), proposed by Pearl Judea in 1987, offer a better solution. BNs comprise directed acyclic graph (DAG), reflecting potential relationships among risk factors, and conditional probability tables (CPT), which demonstrate correlations between variables [10,11]. BNs hold many advantages, one of which relates to its unstrict statistical hypothesis [12]. Moreover, with known nodes, BNs could infer the probability of unknown nodes, flexibly showing the impact of relevant risk factors

on stroke. Accordingly, it allows for complex networks between one disease and its risk factors, overcoming the limitations of traditional Logistic regression model.

BN learning refers to the obtainment of complete BNs by existing information. The construction method consists of parameter learning and structure learning. The former seeks for parameter determination based on a known network structure. This study focuses on structure learning, a more commonly used algorithm, which could be divided into score-based search and constraint-based algorithm [12]. The former aims for the best score-functioning BN structure. However, it's hard to obtain an optimal network structure under a large structure space. The latter boasts a higher learning efficiency and allows for a globally optimal solution, but it also comes with some shortcomings. The first one emphasizes sophisticated judgement of node independence. Also, with more nodes, the independence tests between nodes increase exponentially. The second one relates to unreliable high-order conditional independence test. Due to their limitations, some scholars have proposed a hybrid algorithm, Max-Min Hill-Climbing (MMHC) [13], which combines the advantages of both score-search and constraint-based algorithms and draws great attention from researchers. The MMHC algorithm for BN construction consists of two phases. The first phase is to use the heuristic search algorithm, Max-Min Parents and Children algorithm, to identify candidate parent and child nodes, thereby building a BN framework that detects the network with the highest scores. The second phase is to perform a scoring search to determine the edges and directions of the network structure. As such, MMHC algorithm could reduce the size of the research size and determine the optimal network structure. At present, BN with MMHC algorithm has been employed to discuss the risk factors for chronic obstructive pulmonary disease, hyperlipidemia and other diseases [13], showing great performance.

Yet, data imbalance is widespread in clinical studies, because the number of positives is much smaller than that in the normal population. As mentioned above, there are 298 cases per 100,000 person-years of stroke in rural areas, so a category imbalance is prominent in the stroke dataset. In data-driven algorithms, an imbalanced dataset would contribute to lower model performance, so it's important to balance stroke categories before constructing BN model [14]. It also has been documented that when discussing stroke prediction using data-driven models, an unbalanced data was first handled to report the results [15,16]. Propensity score matching (PSM) [17] has been shown a good approach to reducing the bias due to confounding variables. In this study, some variables, including smoking, alcohol consumption, diet and salt consumption were not well-defined and it's inappropriate to take these variables into analysis. Accordingly, it's good to conduct propensity scores with 1:1 matching to eliminate the influence of these variables and to create balanced stroke categories for better model construction.

This study aimed to employ BN model with MMHC algorithm to explore risk factors for stroke, combined with PSM, with data from stroke screening in Shanxi Province, thus providing a new idea for clinical practice, reducing the prevalence of stroke and improving the quality of life.

2. Materials and methods

2.1. Study participants

From April 2019 to November 2019, a screening program for chronic kidney disease and stroke was conducted in ten rural regions in Shanxi province, i.e., Ningwu county, Yu county, Yangqu county, Lin county, Shouyang county, Zezhou county, Huozhou city, Hejin city, Linyi county, Ruicheng county.

In total, 13,550 villagers participated in the program, and 12,285 were finally enrolled in this study with 5206 men and 7079 women. Informed consent was signed by all study participants and this study was approved by the Ethics Committee of Shanxi Provincial hospital, with reference number 2021213. Inclusion criteria included residents over 40 years old. Exclusion criteria included incomplete recorded data; those less than 40 years old; those unwilling to cooperate and pregnant women with and history of substance abuse.

2.2. Data collection

Data were collected by questionnaire, physical examinations, and laboratory analyses. The questionnaire comprises sociodemographic information (sex, age, annual income, educational levels), stroke family history, and lifestyles (exercise, smoking, alcohol consumption, dietary habits). The questionnaire was conducted online or offline and was completed by the subjects themselves or their families. Physical examinations were conducted by medical staff or trained medical students at Shanxi Provincial hospital. They consist of heigh, weight, and blood pressure which was measured two times and then we calculated the mean value. Body Mass Index (BMI) was calculated as weight in kilograms divided by the square of height in meters. Besides, fasting venous blood was taken from participants for high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC), triglyceride (TG), glycosylated haemoglobin (GHb), fasting plasma glucose (FPG), homocysteine (Hcy). Additionally, morning urine was garnered for α1-Microgloblobumin (α1-MG), urine creatinine (Ucr) and urinary albumin (mAlb), and then we calculated Urinary albumin-to-creatinine ratio (ACR) and α1-MG-to-creatinine ratio (MCR).

2.3. Propensity score matching

Generally, each covariate variable is substituted into the Logistic regression model, and the conditional probability of being sorted into the case group is calculated, thus detecting the control group individuals with similar characteristics to each case [18]. In this study, the population characteristics such as smoking, alcohol consumption, diet, and salt consumption were used as matching variables for their unperfect definition, and the caliper value was set at 0.1, and the calipers were matched with a ratio of 1:1 for 748 stroke cases and 11,537 non-stroke cases in the non-stroke group.

2.4. Bayesian network (BN)

A BN consists of a directed acyclic graph whose nodes represent random variables and edges express dependencies between nodes [19]. If one edge is from variable A to variable B (A→B), it means variable A has a direct influence on variable B or variable A is the risk factor for variable B. Also, we name variable A the parent node of variable B, and variable B the child node of variable A. Another important concept in BN is that each child node has a conditional probability distribution that measures the effects of its predictor variables (parents). This is given by P (Xi| pa (Xi)), where P is the conditional probability, Xi represents each node and pa (Xi) represents the parents of node Xi [20]. BNs use the graphical structure and network parameters to uniquely determine the joint probability distribution on the random variable = x {X1, Xn}, which can be listed as:

$$P(x_1,\ x_2,\ \cdots,\ x_n) = P(x_1)P_{(x_2|x_1)} \cdots P(x_{n|x_1,x_2,\cdots,x_{n-1}}) = \Pi_1^n P(x_i|\pi(x_i)) \quad (1)$$

$\pi(x_i)$ is the set of parent nodes of x_i, $\pi(x_i) \subseteq (x_1, \ldots, x_{i-1})$. When the value of $\pi(x_i)$ is known, x_i is conditionally independent of other variables in $(x_1, \ldots, x_{i-1})$. Figures 1 and 2 are examples of BN. The BN consists of three nodes, including hypertension, cardiovascular disease and depression. The edges points from hypertension to cardiovascular disease and depression. It means that both cardiovascular disease and depression are the child nodes of hypertension, suggesting that hypertension has a direct influence on both cardiovascular disease and depression. Figure 1 shows that the prior probability of cardiovascular disease and depression stands at 0.559 and 0.0815. If one is subjected to hypertension, the probability increases from the prior probability to P (cardiovascular disease|hypertension) = 0.896 and to P (depression|hypertension) = 0.125 (Figure 2).

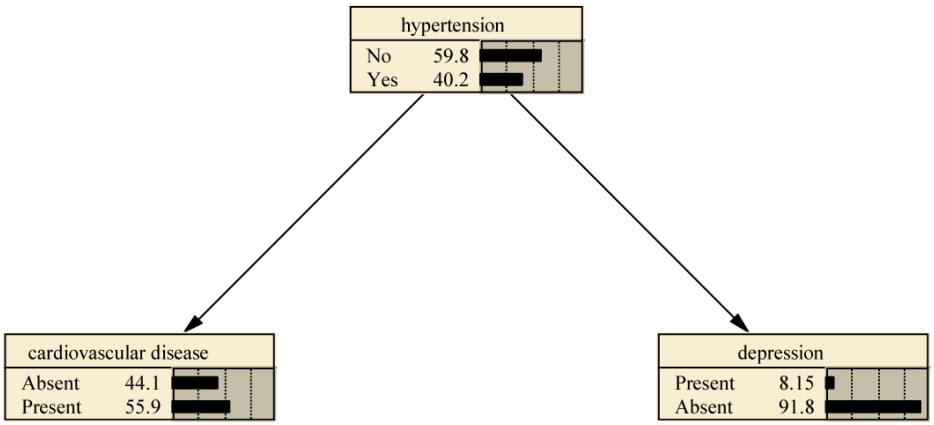

Figure 1. An example of Bayesian network model.

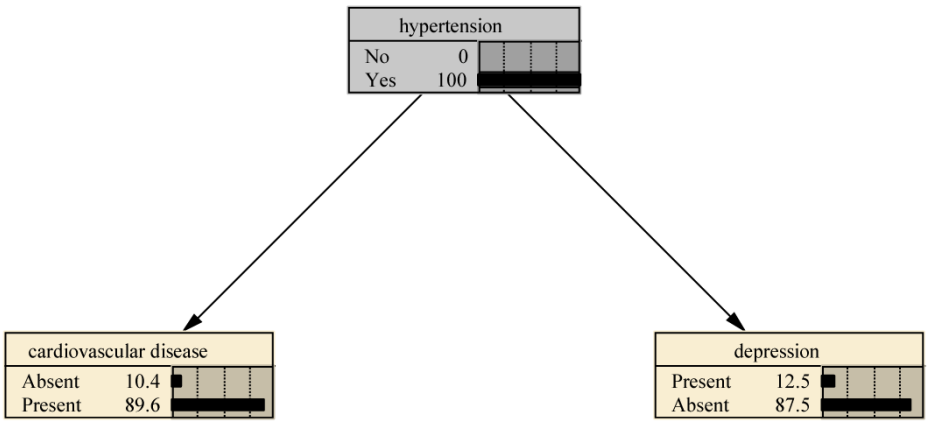

Figure 2. An example of Bayesian reasoning.

2.5. MMHC algorithm

MMHC algorithm is one of the most widely-used hybrid algorithms that could overcome the

drawbacks by combining the advantages of both the constraint-based method and search score method [21]. MMHC algorithm construction comprises two phases. The first one is to employ a heuristic search algorithm, the Max-Min Parents and Children algorithm, to determine candidate parent and child nodes, thus constructing a BN framework and detecting the network with the largest score through increasing, removing, and transforming the direction of the edges in the constraint space through the method of scoring search. The second one is to perform a scoring search to determine the edges of the network structure and the orientation of the edges. It involves a greedy search method, which starts with a blank graph and achieves the highest rated belief network structure by continuously adding edges, subtracting edges and reversing directions to the network [13,21].

2.6. Definitions

Annual income/educational levels, smoking/alcohol consumption status, previous medical history, and lifestyle were obtained from a questionnaire. The categories of annual income were divided into 4 parts, namely, < 5000 Yuan, 5000–10,000 Yuan, 10,000–20,000 Yuan, and > 20,000 Yuan. The categories of education level consisted of ≤ Primary school, ≤ Middle school, ≤ High school, ≥ Bachelor degree. Smoking was divided into Yes or No. Alcohol consumption was defined as Always (over 100 g/time and 3 times/week), Sometimes (< 3 times/week or less than 100 g/time) and Seldom. Exercise was classified into Never or Always (≥ 3 times/weeks and ≥ 30 min/time with intensity over moderate walking). Salt consumption was defined as Light, Balanced, and Salty. Diet was defined as Vegetable, Balanced and Meat.

Under the Chinese Guidelines on Prevention and Treatment of Dyslipidemia in Adults published in 2007 [22], TC ≥ 6.22 mmol/L was defined as hypercholesterolemia; TG ≥ 2.26 mmol/L was defined as hypertriglyceridemia; LDL-C ≥ 4.14 mmol/L was defined as high levels of low-density lipoprotein cholesterol; HDL-C < 1.04 mmol/L was defined as low levels of high-density lipoprotein cholesterol. Hyperhomocysteinemic was defined as Hcy > 15μmol/L [23]. FPG was defined as Normal (< 6.1 mmol/L), Impaired Fasting Glucose (IFG, 6.1~7.0 mmol/L), Hyperglycosemia (> 7.0 mmol/L).

Following the Guidelines for the Prevention and Treatment of Type 2 Diabetes in China published in 2021, GHb was defined as Normal (< 6.5 mmol/L) and Abnormal (≥ 6.5 mmol/L). systolic blood pressure (SBP) was defined as High (≥ 140 mmHg), Normal (120~140 mmHg) and low (< 120 mmHg), and diastolic blood pressure (DBP) was defined as High (≥ 90 mmHg), Normal (80~90 mmHg) and Low (< 80mmHg) [24]. As per the standards established for Chinese by the Department of Disease Control, Ministry of Health [25], BMI classification comprised underweight (< 18.5 kg/m^2), normal (18.5~24.0 kg/m^2), overweight (24.0~28.0 kg/m^2), obesity (≥ 28 kg/m^2). ACR equals to mAlb divided by Ucr multiplied by 8.84; MCR equals α-1 MG divided by Ucr multiplied by 8.84. ACR ≥ 30 mg/g was defined as increased ACR and MCR > 23 mg/g was defined as increased MCR. Stroke is diagnosed with clear imaging evidence (CT, MRI) by a neurological physician.

2.7. Statistical analyses

Categorical variables were expressed as percentage (%) and we used chi-square tests to determine the difference between groups. Afterwards, a multivariate stepwise logistic regression ($\alpha in = 0.05$, $\alpha out = 0.10$) was employed for exploration of the risk factors for increased ACR and increased MCR; the independent variables were those statistically significant variables in Chi-squared test. Significant

risk factors were incorporated to construct the BN model. Statistical description, Chi-squared test, multivariate logistic regression and structure learning of BN were conducted using R studio 4.2.0 (R Development Core Team). $P < 0.05$ was considered statistically significant. The structure learning was realized using mmhc () function in the package "bnlearn". The BN model and BN reasoning were visualized using Netica software (Norsys Sofware Corp., Vancouver, BC, Canada). Besides, the maximum likelihood estimation was employed for CPT.

3. Results

3.1. Baseline characteristics of study population

Before PSM, there are 11,537 non-stroke cases, and 748 cases in this study, and smoking, diet, salt consumption is incomparable ($P < 0.05$). After PSM, there are 748 non-stroke cases and 748 stroke cases, and smoking, diet, salt consumption, and alcohol consumption is comparable ($P > 0.05$), as shown in Table S1. In stroke group, men account for 52.5 and 43.2% of the patients aged 61 years to 70 years. Nearly half of them are less-educated, with 46.9% of them with ≤ primary educational background. Besides, the annual income is not handsome; 53.1% of them are with an income of ＜5k. Additionally, lots of them are not subject to abnormal biochemical parameters; 77.7, 95.3, 96.3, 80.7, 82.9 and 83.6% of the patients have normal TG, TC, LDL, HDL, FPG, and GHb. Notably, 73% of the patients are subjected to abnormal HHcy. More than 50% of them are women (56.6%), and 36.1% are with the age of 51 to 60 years, with nearly half of them (48.5%) with an educational background of ≤ middle school. Most of them had no dyslipidemia, with 83.4, 95.9, 98.4 and 86.4% not subjected to abnormal TG, TC, LDL and HDL. 67% of them were subjected to HHcy, and 90.1% had no family history of stroke; 90.8% had a normal ACR and 87.8% had a normal MCR. More detailed descriptions are listed in Table 1.

3.2. Univariate analysis

We used chi-square tests to explore the differences in each variable between the stroke and non-stroke groups. The results showed that the differences in exercise and TC between the two groups were not statistically significant ($P > 0.05$). Sex, age, education, income, TG, LDL, HDL, FPG, GHb, Hcy, SBP, DBP, BMI, family history, ACR, MCR were statistically significant between the two groups ($P < 0.05$), as reflected in Table 1.

Table 1. baseline characteristics of stroke and non-stroke groups.

Variables	levels	Non-stroke (N = 748)	Stroke (N = 748)	*P*
sex	men	325 (43.4%)	393 (52.5%)	< 0.001
	women	423 (56.6%)	355 (47.5%)	
age	40~	147 (19.7%)	38 (5.1%)	< 0.001
	51~	270 (36.1%)	187 (25%)	
	61~	240 (32.1%)	323 (43.2%)	
	71~	91 (12.2%)	200 (26.7%)	

Continued on next page

Variables	levels	Non-stroke (N = 748)	Stroke (N = 748)	*P*
Educational levels	≤ primary	252 (33.7%)	351 (46.9%)	< 0.001
	≤ middle	363 (48.5%)	335 (44.8%)	
	≤ high	92 (12.3%)	53 (7.1%)	
	≥ bachelor	41 (5.5%)	9 (1.2%)	
income	< 5k	306 (40.9%)	397 (53.1%)	< 0.001
	5–10k	207 (27.7%)	190 (25.4%)	
	10–20k	81 (10.8%)	63 (8.4%)	
	> 20k	154 (20.6%)	98 (13.1%)	
exercise	no	439 (58.7%)	448 (59.9%)	0.674
	yes	309 (41.3%)	300 (40.1%)	
TG	no	624 (83.4%)	581 (77.7%)	0.006
	yes	124 (16.6%)	167 (22.3%)	
TC	no	717 (95.9%)	713 (95.3%)	0.706
	yes	31 (4.1%)	35 (4.7%)	
LDL	no	736 (98.4%)	720 (96.3%)	0.016
	yes	12 (1.6%)	28 (3.7%)	
HDL	no	646 (86.4%)	604 (80.7%)	0.004
	yes	102 (13.6%)	144 (19.3%)	
FPG	normal	686 (91.7%)	620 (82.9%)	< 0.001
	impaired	39 (5.2%)	63 (8.4%)	
	high	23 (3.1%)	65 (8.7%)	
GHb	no	684 (91.4%)	625 (83.6%)	< 0.001
	yes	64 (8.6%)	123 (16.4%)	
Hcy	no	247 (33%)	202 (27%)	0.013
	yes	501 (67%)	546 (73%)	
SBP	low	158 (21.1%)	54 (7.2%)	< 0.001
	normal	326 (43.6%)	257 (34.4%)	
	high	264 (35.3%)	437 (58.4%)	
DBP	low	320 (42.8%)	216 (28.9%)	< 0.001
	normal	267 (35.7%)	288 (38.5%)	
	high	161 (21.5%)	244 (32.6%)	
BMI	underweight	19 (2.5%)	14 (1.9%)	0.009
	normal	330 (44.1%)	274 (36.6%)	
	overweight	299 (40%)	329 (44%)	
	obesity	100 (13.4%)	131 (17.5%)	
Stroke^{f}*	no	674 (90.1%)	628 (84%)	< 0.001
	yes	74 (9.9%)	120 (16%)	
ACR	normal	679 (90.8%)	598 (79.9%)	< 0.001
	increased	69 (9.2%)	150 (20.1%)	
MCR	normal	657 (87.8%)	591 (79%)	< 0.001
	increased	91 (12.2%)	(21%)	

* Stroke^{f}: family history of stroke.

3.3. Multivariate analysis

We conducted a multivariate logistic regression model with stepwise method ($\alpha_{in} = 0.05$, $\alpha_{out} = 0.10$) for risk factors for stroke, with stroke presence as the dependent variables; independent variables were those significantly associated with stroke presence in univariate analysis. The multivariate analysis suggested that stroke was significantly associated with sex (OR 0.718, CI: 0.567–0.91), age (OR 1.833, CI:1.598, 2.102), educational levels (OR 0.764, CI: 0.649, 0.901), HDL (OR 1.508, CI: 1.105, 2.058), FPG (OR 1.413, CI: 1.118, 1.785), SBP (OR 1.657, CI: 1.363, 2.015), DBP (OR 1.239, CI: 1.042, 1.473), family history of stroke (OR 2.048, CI: 1.451, 2.89), ACR (OR 1.865, CI: 1.33, 2.615). Among them family history, ACR, age, SBP constitute the strongest risk factors for stroke, as shown in Table 2.

Table 2. Risk factors of stroke using Logistic regression model.

Variables	B	S.E.	Wald	P	OR (95% C.I.)
sex	0.331	0.121	7.489	0.006	0.718 (0.567, 0.910)
age	0.606	0.070	74.927	< 0.001	1.833 (1.598, 2.102)
education	0.269	0.084	10.276	0.001	0.764 (0.649, 0.901)
HDL	0.411	0.159	6.695	0.010	1.508 (1.105, 2.058)
FPG	0.346	0.119	8.383	0.004	1.413 (1.118, 1.785)
SBP	0.505	0.100	25.646	< 0.001	1.657 (1.363, 2.015)
DBP	0.214	0.088	5.867	0.015	1.239 (1.042, 1.473)
Stroke^{f}*	0.717	0.176	16.629	< 0.001	2.048 (1.451, 2.89)
ACR	0.623	0.172	13.061	< 0.001	1.865 (1.33, 2.615)
Constant	-2.873	0.301	91.110	< 0.001	0.057

* Stroke^{f}: family history of stroke.

3.4. Bayesian networks model

BN was constructed with 10 nodes and 12 directed edges, which is crystal clear in reflecting the risk factors than the Logistic regression model. Directed edges represent probabilistic dependencies between connected nodes. The results suggested that age, FPG, SBP, and stroke family history constitute direct risk factors for stroke, whereas sex, educational levels, HDL, DBP, and ACR represent indirect risk factors for stroke. Besides, the model suggested that age and educational levels are direct risk factors for SBP and FPG is direct risk factors for ACR. Additionally, SBP has a direct influence on DBP and sex is correlated with ACR and HDL (Figure 3).

3.5. Bayesian reasoning

Prior probabilities of the variables are presented in Figure 3. The resulting probabilistic model could quantitatively analyse the influence of these factors on stroke via computing conditional probabilities P (y|xi). From Figure 3, we could learn that the prior probability of stroke stands at 0.50. If one's gender is female, the probability increases from the prior probability to P (stroke|female) = 0.502 (Figure S1). And if the person is subjected to hyperglycemia, the probability rises to P (stroke|female, hyperglycemia) = 0.624 (Figure S2). If the person is over seventy years old, the probability rises to P

(stroke|female, hyperglycemia, 71–91 years) = 0.68 (Figure S3). If he is also subjected to high SBP, the probability amounts to P (stroke|female, hyperglycemia, 71–91 years, high SBP) = 0.725 (Figure S4).

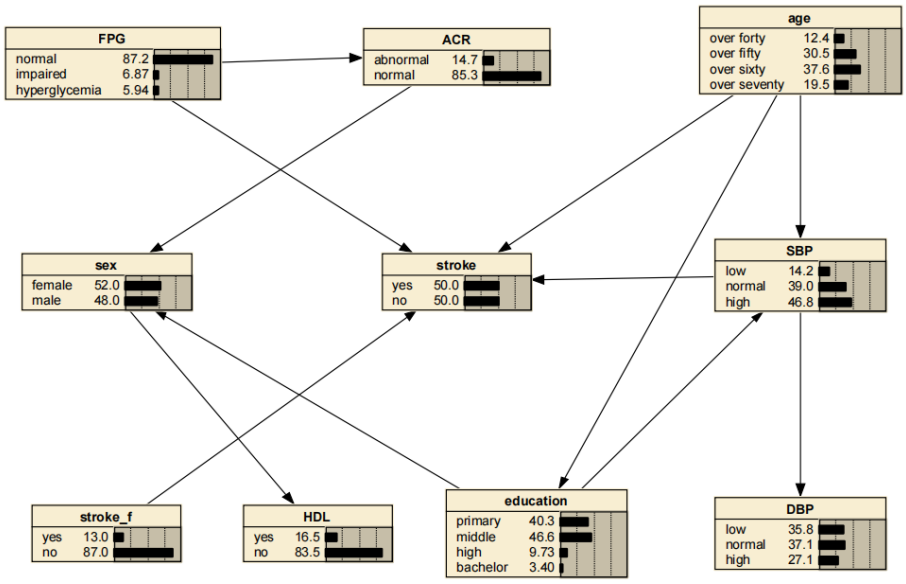

Figure 3. MMHC algorithm to construct stroke BNs and prior probability.

4. Discussion

When exploring risk factors for stroke, past studies often employed logistic regression, which uses probabilities to reflect the strength of the association, however, it cannot elaborate on the overall association between risk factors [21], nor can it detect direct or indirect risk factors. BNs, on the other hand, boast more advantages than the former in constructing risk factor models [26]. Firstly, BNs do not require any prior assumptions. Secondly, the model can integrate different variables and analyze their relative importance [27]. Therefore, BNs have been favoured by many clinical researchers in recent years. Studies have shown that BNs, as risk assessment tools for large clinical datasets, can quantitatively identify indicators important for predicting specific histopathological diagnoses, and prognosis and identifying risk factors for diseases that support medical decision-making [12]. Of note, the more variables there are when constructing BNs, the more complex the network is, so BNs should be constructed based on univariate and multivariate analysis.

In this study, logistic regression model shows that sex, age, educational levels, HDL, FPG, SBP, DBP, family history of stroke, ACR represent risk factors for stroke. Yet, BNs with MMHC algorithm suggest that age, FPF, SBP and stroke family history represent direct risk factors for stroke, and sex, age, educational background, HDL, DBP and ACR constitute indirect risk factors for stroke. The risk factors identified in our findings are generally consistent with previous studies [28–31]. Besides, BN with MMHC algorithm shows that education levels could be an indirect risk factor for stroke through SBP. Age could be a direct and indirect risk factor for stroke through SBP, indicating its ability to explore the intermediate linkages between associated factors and stroke, and its suitability to detect variables related to stroke.

To the best of our knowledge, our study is the first to apply BNs to the risk factors for stroke. It

not only reveals the risk factors for stroke, but also determines the direct and indirect effects of their effects on stroke and elucidates their complex network relationships. BN with MMHC has an advantage over traditional logistic regression model in risk factor analysis for stroke. The first one concerns variable dependency. As BNs with MMHC algorithm shows sex is related to HDL, ACR and education levels, suggesting that there is an intercorrelation between risk factors for stroke. Logistic regression model is constructed under the premise that variables should be inter-independent, which fails to fully exploit data information and truly reflect the impact on stroke. As such, it's unable to provide a scientific idea for the prevention and control of stroke. BN with MMHC algorithm is a data-driven model constructed on disease-related knowledge, having no strict requirements for the data distribution. Therefore, it facilitates discovering potential unobvious but important data information, offering a scientific foundation for stroke evaluation, prediction and prevention.

The second one relates to the interaction between variables. Logistic regression could only suggest those risk factors for stroke, while BN with MMHC algorithm allows for further description of how these risk factors are interrelated and affect the occurrence of stroke through a graphical approach, which holds certain significance to offer research clues and potential risk factors for stroke. When there is an interaction between variables, extra interaction analysis combined with logistic regression is needed. Yet, BN with MMHC algorithm facilitates direct detection of interaction between risk factors, which helps to comprehensively explore the internal relationship between risk factors.

The incidence of hypertension, atherosclerosis, intimal thickening and narrowing of the arteries increases progressively with age, and metabolic disturbances become more pronounced. Primary cerebral haemorrhage is most often seen in patients of advanced age, as it is mainly associated with hypertension and atherosclerosis [32,33].

Atherosclerotic thrombotic cerebral infarction is most common in older people. For elderly patients with obvious hypertension and atherosclerosis, when blood pressure fluctuates greatly, there is a sudden decrease in blood volume such as a lot of sweating, surgical bleeding, septic shock, and severe diarrhoea leading to low blood pressure, and cerebral infarction can be triggered again [34].

In patients with poor blood glucose management, increased blood viscosity, prolonged hypercoagulability, increased likelihood of thrombosis, abnormal function of platelets and fibrinogen, and prolonged activation time of coagulation factors in the blood may lead to a significant increase in the recurrence rate of elderly stroke patients. Hyperglycemia causes endothelial cell damage and promotes the expression of inflammatory factors, which is closely related to the occurrence and development of atherosclerosis [35].

Haemodynamic changes have a direct impact on stroke. Sudden fluctuations in blood pressure can cause plaque on the intima to rupture and components of the blood [36] such as platelets and fibrin to adhere, aggregate and deposit to form thrombi, leading to cerebral infarction. Reduced blood flow can also increase the risk of plaque formation in the cerebral arteries. Plaque in the cerebral arteries can also cause significant narrowing or occlusion of the lumen itself, resulting in reduced blood pressure, slower blood flow and increased blood viscosity in the perfused area, which in turn can reduce blood supply to the local brain area or promote local thrombosis and symptoms of cerebral infarction [36]. Haemorrhagic strokes are represented by cerebral haemorrhage, which is caused by abnormal fluctuations in blood pressure, leading to rupture of blood vessels in the brain parenchyma.

Genetic factors are directly related to stroke. For example, subcortical atherosclerotic encephalopathy, a disease that accompanies white matter lesions, is an autosomal dominant cerebrovascular disease [37]. Generally, this hereditary disease is a recurrent stroke occurrence. Or

polygenic genetic disorders, such as patients with familial hypertension, do not necessarily lead to stroke but in combination with environmental factors such as lifestyle habits, the risk of stroke is greatly increased.

This study also has some limitations. First, the directed edges in BNs cannot represent causal relationships between connected nodes, but rather probabilistic dependencies. Second, since this study was based on an opportunistic screening program and the locations were mainly located in southern or central Shanxi Province, this study may be subject to selection bias. Third, some of the data were collected through questionnaires, so recall bias may exist. Third, we did not classify stroke into ischemic stroke and hemorrhagic stroke. Our ongoing work is to focus on different types of stroke. Finally, since this study focused on BNs with MMHC algorithm, we didn't make a comparison with other hybrid algorithms, which will also be the focus of our future work.

5. Conclusions

Risk factor detection is of great significance for disease prevention. Our study suggested that BN with MMHC algorithm not only could achieve a complex network relationship between risk factors and stroke, but also makes risk prediction for stroke possible, providing a scientific idea for stroke control and treatment, helping lower the prevalence of stroke. Specific findings could be listed below：

1) Logistic regression model demonstrated that stroke was significantly associated with sex, age, educational levels, HDL, FPG, SBP, DBP, family history of stroke, and ACR.

2) The BN model for stroke was constructed with 10 nodes and 12 directed edges. Age, FPG, SBP, and family history of stroke represent direct risk factors for stroke, whereas sex, educational levels, HDL, DBP, and ACR constitute indirect risk factors for stroke.

3) BN with MMHC algorithm could achieve probability inference of unknown nodes through known nodes, flexibly demonstrating the impact of one risk factor on stroke.

4) BN with MMHC algorithm outperforms traditional logistic regression model, which boasts great prospects in clinical practice.

Acknowledgements

This work was supported by the Key Laboratory Project of Shanxi Province(201805D111020) and the Key Laboratory Construction Plan Project of Shanxi Provincial Health Commission(2020SYS01). We also appreciate all the authors and patients participating in this study.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. C. M. Stinear, C. E. Lang, S. Zeiler, W. D. Byblow, Advances and challenges in stroke rehabilitation, *Lancet Neurol*., **19** (2020), 348–360. https://doi.org/10.1016/S1474-4422(19)30415-6

2. C. Iadecola, M. S. Buckwalter, J. Anrather, Immune responses to stroke: mechanisms, modulation, and therapeutic potential, *J. clin. invest*., **130** (2020), 2777–2788. https://doi.org/10.1172/JCI135530
3. S. Wu, B. Wu, M. Liu, Z. Chen, W. Wang, C. S. Anderson, et al., Stroke in China: advances and challenges in epidemiology, prevention, and management, *Lancet Neurol*., **18** (2019), 394–405. https://doi.org/10.1016/S1474-4422(18)30500-3
4. W. Wang, B. Jiang, H. Sun, X. Ru, D. Sun, L. Wang, et al., Prevalence, incidence, and mortality of stroke in China: results from a nationwide population-based survey of 480 687 adults, *Circulation*, **135** (2017), 759–771. https://doi.org/10.1161/CIRCULATIONAHA.116.025250
5. X. Xia, W. Yue, B. Chao, M. Li, L. Cao, L. Wang, et al., Prevalence and risk factors of stroke in the elderly in Northern China: data from the National Stroke Screening Survey, *J. Neurol.*, **266** (2019), 1449–1458. https://doi.org/10.1007/s00415-019-09281-5
6. Y. Wu, Y. Fang, Stroke prediction with machine learning methods among older Chinese, *Int. J. Environ. Res. Public Health*, **17** (2020), 1828. https://doi.org/10.3390/ijerph17061828
7. A. Aigner, U. Grittner, A. Rolfs, B. Norrving, B. Siegerink, M. A. Busch, Contribution of established stroke risk factors to the burden of stroke in young adults, *Stroke*, **48** (2017), 1744–1751. https://doi.org/10.1161/STROKEAHA.117.016599
8. Y. Dong, W. Cao, X. Cheng, K. Fang, X. Zhang, Y. Gu, et al., Risk factors and stroke characteristic in patients with postoperative strokes, *J. Stroke Cerebrovasc. Dis.*, **26** (2017), 1635–1640. https://doi.org/10.1016/j.jstrokecerebrovasdis.2016.12.017
9. Z. Wei, X. L. Zhang, H. X. Rao, H. F. Wang, X. Wang, L. X. Qiu, Using the Tabu-search-algorithm-based Bayesian network to analyze the risk factors of coronary heart diseases, *Chin. J. Epidemiol*., **37** (2016), 895–899. https://doi.org/10.3760/cma.j.issn.0254-6450.2016.06.031
10. S. J. Moe, J. F. Carriger, M. Glendell, Increased use of bayesian network models has improved environmental risk assessments, *Integr. Environ. Assess. Manage*., **17** (2021), 53–61. https://doi.org/10.1002/ieam.4369
11. A. Frolova, B. Wilczyński, Distributed Bayesian networks reconstruction on the whole genome scale, *PeerJ*, **6 (**2018), e5692. https://doi.org/10.7717/peerj.5692
12. J. Pan, H. Rao, X. Zhang, W. Li, Z. Wei, Z. Zhang, et al., Application of a Tabu search-based Bayesian network in identifying factors related to hypertension, *Medicine*, **98** (2019), e16058. https://doi.org/10.1097/MD.0000000000016058
13. D. Quan, J. Ren, H. Ren, L. Linghu, X. Wang, M. Li, et al., Exploring influencing factors of chronic obstructive pulmonary disease based on elastic net and Bayesian network, *Sci. Rep.*, **12** (2022), 7563. https://doi.org/10.1038/s41598-022-11125-8
14. Z. Xu, D. Shen, T. Nie, Y. Kou, A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data, *J. Biomed. Inf.*, **107** (2020), 103465. https://doi.org/10.1016/j.jbi.2020.103465
15. M. S. Pathan, A Nag, M. M. Pathan, S. Dev, Analyzing the impact of feature selection on the accuracy of heart disease prediction, *Healthcare Anal.*, **2** (2022), 100060. https://doi.org/10.1016/j.health.2022.100060
16. S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, D. John, A predictive analytics approach for stroke prediction using machine learning and neural networks, *Healthcare Anal.*, **2** (2022), 100032. https://doi.org/10.1016/j.health.2022.100032

17. L. T. Kane, T. Fang, M. S. Galetta, D. K. C. Goyal, K. J. Nicholson, C. K. Kepler, et al., Propensity score matching: a statistical method, *Clin. Spine Surg.*, **33** (2020), 120–122. https://doi.org/10.1097/BSD.0000000000000932
18. J. Liang, Z. Hu, C. Zhan, Q. Wang, Using propensity score matching to balance the baseline characteristics, *J. Thorac. Oncol.*, **16** (2021), E45–E46. https://doi.org/10.1016/j.jtho.2020.11.030
19. E. Park, H. J. Chang, H. S. Nam, A Bayesian network model for predicting post-stroke outcomes with available risk factors, *Front. Neurol.*, **9** (2018), 699. https://doi.org/10.3389/fneur.2018.00699
20. D. E. da Cunha Leme, The use of Bayesian network models to identify factors related to frailty phenotype and health outcomes in middle-aged and older persons, *Arch. Gerontol. Geriatr.*, **92** (2021), 104212. https://doi.org/10.1016/j.archger.2020.104212
21. X. Wang, J. Pan, Z. Ren, M. Zhai, Z. Zhang, H. Ren, et al., Application of a novel hybrid algorithm of Bayesian network in the study of hyperlipidemia related factors: a cross-sectional study, *BMC Public Health*, **21** (2021), 1375. https://doi.org/10.1186/s12889-021-11412-5
22. Y. Huang, L. Gao, X. Xie, S. C. Tan, Epidemiology of dyslipidemia in Chinese adults: meta-analysis of prevalence, awareness, treatment, and control, *Popul. Health Metrics*, **12** (2014), 1–9. https://doi.org/10.1186/s12963-014-0028-7
23. L. P. Zhao, T. You, S. P. Chan, J. C. Chen, W. T. Xu, Adropin is associated with hyperhomocysteine and coronary atherosclerosis, *Exp. Ther. Med.*, **11** (2016), 1065–1670. https://doi.org/10.3892/etm.2015.2954
24. Z. Wang, Z. Chen, L. Zhang, X. Wang, G. Hao, Z. Zhang, et al., Status of hypertension in China: results from the China hypertension survey, 2012–2015, *Circulation*. **137** (2018):2344–2356. https://doi.org/10.1161/CIRCULATIONAHA.117.032380
25. N. Shi, K. Liu, Y. Fan, L. Yang, S. Zhang, X. Li, et al., The association between obesity and risk of acute kidney injury after cardiac surgery, *Front. Endocrinol.*, **11** (2020), 534294. https://doi.org/10.3389/fendo.2020.534294
26. P. Arora, D. Boyne, J. J. Slater, A. Gupta, D. R. Brenner, M. J. Druzdzel, Bayesian networks for risk prediction using real-world data: a tool for precision medicine, *Value health*, **22** (2019):439–445. https://doi.org/10.1016/j.jval.2019.01.006
27. Y. Dimitrov, M. Ducher, M. Kribs, G. Laurent, S. Richter, J. P. Fauvel, Variables linked to hepatitis B vaccination success in non-dialyzed chronic kidney disease patients: use of a bayesian model, *Nephrol. Ther.*, **15** (2019), 215–219. https://doi.org/10.1016/j.nephro.2019.02.010
28. C. S. Anderson, Progress-defining risk factors for stroke prevention, *Cerebrovasc. Dis.*, **50** (2021), 615–616. https://doi.org/10.1159/000516996
29. W. Qi, J. Ma, T. Guan, D. Zhao, A. Abu-Hanna, M. Schut, et al., Risk factors for incident stroke and its subtypes in China: a prospective study, *J. Am. Heart Assoc.*, **9** (2020), e016352. https://doi.org/10.1161/JAHA.120.016352
30. C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, D. John, Predicting stroke from electronic health records, in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Berlin, Germany, (2019), 5704–5707. https://doi.org/10.1109/EMBC.2019.8857234
31. M. S. Pathan, Z. Jianbiao, D. John, A. Nag, S. Dev. Identifying stroke indicators using rough sets, *IEEE Access*, **8** (2020), 210318–210327, https://doi.org/10.1109/ACCESS.2020.3039439

32. M. N. Cocchi, J. A. Edlow, Managing hypertension in patients with acute stroke, *Ann. Emerg. Med*. **75** (2020), 767–771. https://doi.org/10.1016/j.annemergmed.2019.09.015
33. Y. C. Cheng, J. M. Sheen, W. L. Hu, Y. C. Hung. Polyphenols and oxidative stress in atherosclerosis-related ischemic heart disease and stroke, *Oxid. Med. Cell. Longevity*, **2017** (2017), 8526438. https://doi.org/10.1155/2017/8526438
34. S. N. Bhupathiraju, F. B. Hu, Epidemiology of obesity and diabetes and their cardiovascular complications, *Circ. Res.*, **118** (2016), 1723–1735. https://doi.org/10.1161/CIRCRESAHA.115.306825
35. F. Denorme, I. Portier, Y. Kosaka, R. A. Campbell, Hyperglycemia exacerbates ischemic stroke outcome independent of platelet glucose uptake, *J. Thromb. Haemostasis*, **19** (2021), 536–546. https://doi.org/10.1111/jth.15154
36. S. L. Stevens, S. Wood, C. Koshiaris, K. Law, P. Glasziou, R. J. Stevens, et al., Blood pressure variability and cardiovascular disease: systematic review and meta-analysis, *BMJ*, **354** (2016), i4098. https://doi.org/10.1136/bmj.i4098
37. X. Zheng, N. Zeng, A. Wang, Z. Zhu, H. Peng, C. Zhong, et al., Family history of stroke and death or vascular events within one year after ischemic stroke, *Neurol. Res.*, **41** (2019), 466–472. https://doi.org/10.1080/01616412.2019.1577342

AIMS Press