Mathematical Biosciences and Engineering

MBE, 20(9): 16194–16211.
DOI: 10.3934/mbe.2023723
Received: 02 June 2023
Revised: 26 July 2023
Accepted: 27 July 2023
Published: 10 August 2023

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

Research article

Using Bayesian networks with tabu algorithm to explore factors related to chronic kidney disease with mental illness: A cross-sectional study

Xiaoli Yuan[1], Wenzhu Song[2], Yaheng Li[3], Qili Wang[2], Jianbo Qing[1], Wenqiang Zhi[1], Huimin Han[1], Zhiqi Qin[4], Hao Gong[4], Guohua Hou[5] and Yafeng Li[1,5,6,7,*]

[1] Department of Nephrology, Shanxi Provincial People's Hospital (Fifth Hospital) of Shanxi Medical University, Taiyuan 030012, China
[2] School of Public Health, Shanxi Medical University, No.56 Xinjian South Road, Taiyuan 030001, China
[3] Shanxi Provincial Key Laboratory of Kidney Disease, Taiyuan 030012, China
[4] Department of Biochemistry & Molecular Biology, Shanxi Medical University, Taiyuan, Shanxi, China
[5] Department of Nephrology, Hejin People's hospital, Yuncheng 043300, China
[6] Core Laboratory, Shanxi Provincial People's Hospital (Fifth Hospital) of Shanxi Medical University, Taiyuan 030012, China
[7] Academy of Microbial Ecology, Shanxi Medical University, Taiyuan 030012, China

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

Abstract: While Bayesian networks (BNs) offer a promising approach to discussing factors related to many diseases, little attention has been poured into chronic kidney disease with mental illness (KDMI) using BNs. This study aimed to explore the complex network relationships between KDMI and its related factors and to apply Bayesian reasoning for KDMI, providing a scientific reference for its prevention and treatment. Data was downloaded from the online open database of CHARLS 2018, a population-based longitudinal survey. Missing values were first imputed using Random Forest, followed by propensity score matching (PSM) for class balancing regarding KDMI. Elastic Net was then employed for variable selection from 18 variables. Afterwards, the remaining variables were included in BNs model construction. Structural learning of BNs was achieved using tabu algorithm and the parameter learning was conducted using maximum likelihood estimation. After PSM, 427 non-KDMI cases and 427 KDMI cases were included in this study. Elastic Net identified 11 variables significantly associated with KDMI. The BNs model comprised 12 nodes and 24 directed edges. The results suggested that diabetes, physical activity, education levels, sleep duration, social activity, self-report on health and asset were directly related factors for KDMI, whereas sex, age, residence and

Internet access represented indirect factors for KDMI. BN model not only allows for the exploration of complex network relationships between related factors and KDMI, but also could enable KDMI risk prediction through Bayesian reasoning. This study suggests that BNs model holds great prospects in risk factor detection for KDMI.

Keywords: Bayesian networks; chronic kidney disease with mental illness (KDMI); tabu algorithm; related factors; model construction; elastic net; propensity score matching

1. Introduction

Chronic kidney disease (CKD) represents a public health issue characterized by progressive loss of renal function, affecting about 10.8% of the population in China [1]. The prevalence of end-stage kidney disease (non-dialysis) and dialysis account for 23% and 46%, respectively [2–4]. Patients with CKD often experience symptoms such as fatigue, sleep disorders and anxiety. CKD has been linked to a higher prevalence of depression compared to other chronic diseases [5] and individuals with severe mental illness are more susceptible to developing CKD [6]. The coexistence of these two could be responsible for poorer medical outcomes, social dysfunction and neurocognitive disorders [7]. Additionally, anxiety and feelings of desperation regarding prognosis and treatment effectiveness may contribute to a lower quality of life and increased social medical burden for individuals with CKD [8]. Besides, life expectancy for those with severe mental illness is 10–20 years shorter than the general population [9], resulting in increased hospitalization rates, longer hospital stays and higher mortality risk [10–12].

In our previous work [13], we explored factors related to multimorbidity using data from the online open database, The China health and retirement longitudinal studies (CHARLS) 2018. However, the definition of multimorbidity as a person subjected to two or more chronic disease conditions from 14 chronic diseases may not be sufficiently reliable. As such, in this study, we focused on the complex relationship between CKD and mental illness, defined as KDMI, aiming to identify associated factors to reduce its prevalence, improve the long-term prognosis of CKD and enhance patients' quality of life.

Traditional logistic regression (LR) model [14,15] has been commonly employed to explore the factors related to KDMI. Yet, the model comes with some defects. The first one concerns variable independency [16]. Clinical diseases often result from intricate interactions between multiple factors, leading to complex interrelationships between variables. Consequently, multiple collinearities may emerge when modeling LR, making it challenging to accurately reveal the relationships between variables. The second one pertains to its inability to facilitate sequential prediction [17]. In most cases, a patient's clinical parameters are incomplete, often due to intentional information concealment, ranging from lifestyle to family medical history, contributing to incomplete data. This incomplete data hampers the feasibility of making probabilistic predictions using LR. Hence, a more robust model is warranted.

Due to advances in data modelling tools, Bayesian networks (BNs) offers a solution. Proposed by Judea Pearl in the 1980s, BNs is a probabilistic graphical model that has become a hot topic of research in recent years. It consists of a directed acyclic graph (DAG) [18], which includes nodes and edges. Nodes represent variables and edges between nodes represent the mutual relationship between nodes

(from the parent node to its child nodes). Conditional probabilities are used to represent the strength of relationships between variables and their dependencies. Currently, BNs has been utilized for factors related to stroke [19], multimorbidity [13], chronic obstructive pulmonary disease [20], hypertension [21] and other diseases.

BNs learning [22] involves obtaining the full BN using existing information, comprising structure learning and parameter learning. The former could be made possible using a constraint-based (CB) algorithm and a score-based search (SS) algorithm. CB is subject to the sophisticated judgement of node independence. Also, with more nodes, the independence tests between nodes increase exponentially. SS could obtain the best score-function BN structure, among which, the tabu-search algorithm, a metaheuristic approach proposed by Glover, has demonstrated superior performance and stands out as one of the most efficient optimization techniques by employing adaptive memory to escape local search and find the global optimum [23].

Due to the multitude of KDMI-related factors, incorporating all of them into the construction of Bayesian networks would result in a highly complex network. Moreover, including weakly-related variables may lead to a less accurate network. Therefore, it is advisable to conduct variable selection before constructing Bayesian networks. Elastic net (EN), proposed by Zou et al. [24] represents a good approach. EN is a well-established regression model combining the advantages of ridge regression and least absolute shrinkage and selection operator (LASSO) regression. Ridge regression contracts the regression coefficients by adding the L2 norm of the regression coefficients, thereby increasing the stability of the estimates. LASSO regression contracts by adding the regression coefficient of the L1 norm regression coefficient. EN introduces both the L1 penalty and the L2 penalty into the minimization process of the objective function and maintains the regular properties of ridge regression while obtaining the sparse coefficient. Consequently, EN serves as an effective method for variable selection in this context.

Another challenge in clinical studies is data imbalance, as the number of positives is much smaller than the number of negatives, translating into a lower model performance in data-driven algorithms [25]. In this study, we addressed data imbalance by utilizing propensity score matching (PSM), a statistical method that mitigates bias in observational studies due to confounding factors [26]. By employing PSM, we aimed to balance the KDMI categories and remove the impact of other chronic diseases, such as cancer, chronic lung disease, liver disease, heart disease, stroke, stomach or digestive system disease, arthritis and memory-related disease, thus achieving a more balanced dataset for constructing the BNs model.

In this study, we first employed PSM to create a balanced dataset for KDMI, after which, EN was used to make a variable selection for stronger variables related to KDMI. Afterwards, BNs with tabu-search algorithm was utilized to discuss factors related to KDMI, thereby offering a scientific basis for KDMI prevention and treatment, lowering the prevalence of KDMI and reducing the social medical burden.

2. Materials and methods

2.1. Study participants

CHARLS is an ongoing longitudinal survey that aims to investigate the social, economic and health conditions of middle-aged and older individuals aged ≥ 45 years old in China. The baseline

survey commenced in 2011 and is followed up every two years, covering 150 districts and 450 urban and rural communities across China, with approximately 10,000 households and 17,000 people, providing a comprehensive picture of the collective situation of middle-aged and older people in China [27].

For this study, we downloaded data from the fourth wave, which was published on September 23, 2018 (http://charls.pku.edu.cn/). All respondents signed informed consent and all CHARLS waves were ethically recognized by the Institutional Review Committee of Peking University. Nine provinces were randomly selected using multi-stage stratified group random sampling in the east, center and west of China.

A questionnaire survey was used to gather information related to family transfer, family information, work retirement, pension, and household income of the population participating in the survey in 2018. Some variables contained missing values, which were addressed using Random Forest. All experiments and methods were performed in accordance with the relevant guidelines and regulations.

2.2. Variable definition

1) General information: Age is divided into < 55 years old, 55–65 years old, 65–75 years old and ≥ 75 years old. Marital status is classified into married, divorced, widowed and never married. Educational background comprises incomplete primary school (≤ primary school), primary school/junior high school (≤ high school), high school/secondary school/junior college (< college), undergraduate and above (≥ college). The residence is classified into town, combination zone between urban and rural areas (boundary), village and special area. Smoking and alcohol consumption is defined as no or yes. Sleep time is divided into ≤ 5 hours, 5–6 hours, 6–7 hours, 7–8 hours and ≥ 8 hours. Sleep duration is classified into ≤ 5 hours, 5–6 hours, 6–7 hours, 7–8 hours and ≥ 8 hours. Asset is defined as ≤ 3000 yuan, ≤ 8000 yuan or > 8000 yuan. Self-health report comprises very poor, poor, fair, good and very good. Internet access is defined as no or yes.

2) Activity of daily living assessment: it consists of basic activities of daily living (BADL) and instrumental activities of daily living (IADL). The former comprises six parameters, including dressing, bathing, eating, going to bed, going to the toilet and controlling defecation. The latter one consists of the following six indicators, doing housework, preparing meals, shopping, managing money and taking their own medicine. Impaired BADL or impaired IADL are defined as the inability to complete any of these parameters [28].

3) Physical activity: The international physical activity questionnaire (IPAQ) [29] was employed for physical activity of the participants. Energy expenditure of physical activity was calculated as follows: exercise intensity assigned to this physical activity × weekly frequency (d/W) × time per day (min/d). The sum of energy expenditure of the three intensity types was the total physical activity expenditure of a week. Vigorous intensity physical activity was assigned 8.0, moderate intensity 4.0 and walking 3.3. Physical activity was divided into three mutually exclusive groups following the calculated physical activity energy expenditure: low (< 600 MET-min/week), moderate (600–3000 MET-min /week) and high (≥ 3000 MET-min/week).

The CHARLS database collected doctor-diagnosed chronic diseases, including 14 types of diseases which were conducted by asking "Have you been diagnosed by a doctor with any of the following diseases", including hypertension, dyslipidemia, diabetes, cancer, chronic lung disease, liver disease, heart disease, stroke, kidney disease, stomach or digestive system disease, emotional or

psychiatric problems, memory-related diseases, arthritis or rheumatism or asthma. In this study, kidney disease in combination with emotional or psychiatric problems was defined as the response variable.

2.3. Propensity Score Matching (PSM)

PSM [30] is a widespread and effective approach to handling confounders in observational studies. Define the observations $S = \{(X_i, T_i, Y_i); i = 1,2, \cdots, n\}$, where X, T and Y are the baseline covariates, dichotomous treatment and outcome variables, respectively. Under the consumption of stable unit treatment value, negligibility and positivity, the causal effect of an individual is defined as $\tau_i = Y_i(1) - Y_i(0)$, where $Y_i(1)$ and $Y_i(0)$ denote the potential outcome of individual i receiving treatment and non-treatment, respectively. The mean causal effect of the treatment group of interest in practical studies can be expressed symbolically as $\mathrm{ATT} = E[Y_i(1) - Y_i(0)|T_i = 1]$ and in estimating ATT, $E[Y_i(1)|T_i = 1]$ is easily obtained from the treatment group observations, while $E[Y_i(0)|T_i = 1]$ is the counterfactual outcome and is not observable. Rosenbaum and Rubin et al. proposed that the counterfactual mean $E[Y_i(0)|T_i = 1]$ in observational studies can be identified by the following equation.

$$E[Y_i(0)|T_i = 1] = \int E[Y|p(X) = \rho, T_i = 0] f_{p|T_i=1}(\rho) d\rho$$

where $f_{p|T_i=1}$ is the distribution of propensity scores $\mathrm{p(X)} = \mathrm{pr(T} = 1|X_i)$ in the treatment group. PS is traditionally estimated by logistic or probit regression models. In this study, other chronic diseases including cancer, chronic lung disease, liver disease, heart disease, stroke, stomach or digestive system disease, arthritis, memory-related disease and asthma were used as matching variables for their potential influence on KDMI, the caliper value was set at 0.1 and the calipers were matched with a ratio of 1:1 for 427 KDMI cases and 19,325 non-KDMI cases.

2.4. Elastic net

In 2005, Zou and Hastie [31] introduced the elastic net (EN) penalty model, which addresses the limitations of LASSO and ridge regression. LASSO regression lacks consideration of associations between features and is not suitable for dealing with multicollinearity, while ridge regression cannot yield predictors with actual coefficients of zero, hindering effective model selection. The EN penalty model represents a convex combination of LASSO and ridge regression and its estimates can be expressed as follows:

$$\widehat{\mathrm{B}} = \arg \min_{\beta^2} \| Y - X\beta \|_2^2 + \lambda \left(\alpha \| B \|_1 + \frac{(\vdash \alpha)}{2} \| B \|_2 \right) \quad (1)$$

Here, λ denotes the penalty coefficient and β represents the regression coefficient. The value of α lies within the range of 0 to 1, enabling adjustment of the penalty with λ. When α = 1, the EN model is equivalent to LASSO regression, while α = 0 renders it equivalent to ridge regression. By combining both LASSO and ridge components, the EN regression offers a more balanced approach, allowing for effective feature selection through cross-validation and compensating for the effects of correlation between observed variables. As a result, the EN regression facilitates the identification of an ideal sparse model, which is essential in high-dimensional data settings.

2.5. Bayesian Networks (BNs)

BNs [32] is a directed acyclic graph (DAG) proposed by Judea Pearl in 1988. It consists of nodes representing variables and directed edges connecting them. Nodes represent random variables and direct edges between nodes represent the interrelationship between nodes (from the parent node to its child nodes). It uses a conditional probability distribution table (CPT) to express the strength of the relationship quantitatively. When this is no parental node, it was expressed with prior probabilities.

Thus, BNs use the graphical structure and network parameters to uniquely determine the joint probability distribution on the random variable = $x\{\mathrm{X1}, \mathrm{Xn}\}$, which can be listed as:

$$\begin{aligned} \mathrm{P}(x_1, x_2, \ldots, x_n) &= \mathrm{P}(x_1)\mathrm{P}(x_2|x_1)\cdots \mathrm{P}(x_n|x_1, x_2, \ldots, x_{n-1}) \\ &= \Pi_1^{\mathrm{n}}\mathrm{P}\big(x_i\big|\pi(x_i)\big) \end{aligned} \tag{2}$$

2.6. Tabu-search algorithm

Tabu search (TS), initially proposed by Professor Fred Glover in 1986 [33], is an intelligent global optimization algorithm widely used for solving complex optimization problems. Inspired by human memory processes, TS efficiently explores the search space to find near-optimal solutions. The algorithm employs a *taboo table* to enhance the search process. This table keeps track of recently explored solutions and prohibits revisiting them in the immediate future, preventing the algorithm from getting stuck in local optima and repetitive cycles. By avoiding redundant searches, TS can more effectively navigate the search space.

Additionally, the taboo table allows the algorithm to make adaptive decisions during the search. It can temporarily *taboo* certain moves that lead to suboptimal solutions, thereby promoting diversification and exploration of different regions. On the other hand, the algorithm retains some promising solutions even in restricted regions, ensuring a balance between exploration and exploitation. As a result, TS demonstrates robustness and the ability to escape local optima, making it particularly effective for optimizing complex problems [34].

2.7. Statistical analysis

Qualitative data are described as percentages (%) and comparisons were made using chi-square tests. EN was achieved using package Glmnet and the structure learning of the BNs was carried out using tabu function in the package bnlearn in R studio (4.2.0). The parameter learning of BNs was achieved using maximum likelihood estimation. Bayesian reasoning and conditional probability distribution tables were plotted using Netica software. $P < 0.05$ was considered statistically significant.

3. Results

3.1. Baseline characteristics of two groups

Before PSM, there were 19,325 cases without KDMI and 427 cases with KDMI. The nine variables, cancer, chronic lung disease, liver disease, heart disease, stroke, stomach or digestive system disease, arthritis, memory-related disease and asthma were statistically significant ($P < 0.05$). After

PSM, the differences between them were statistically insignificant ($P > 0.05$), as shown in Table 1. After PSM, 427 cases with and without KDMI were employed for analysis.

Table 1. Comparisons of potential confounders before and after PSM.

	Before PSM			After PSM		
Variables	Without KDD (19,325)	With KDD (427)	*P*	Without KDD (19,325)	With KDD (427)	*P*
cancer	19,076 (98.7)	416 (97.4)	0.036	416 (97.4%)	416 (97.4%)	1.000
	249 (1.3)	11 (2.6)		11 (2.6%)	11 (2.6%)	
chronic lung disease	18,305 (94.7)	363 (85.0)	<0.001	361 (84.5%)	363 (85%)	0.924
	1020 (5.3)	64 (15.0)		66 (15.5%)	64 (15%)	
liver disease	18,726 (96.9)	375 (87.8)	<0.001	376 (88.1%)	375 (87.8%)	1.000
	599 (3.1)	52 (12.2)		51 (11.9%)	52 (12.2%)	
heart disease	17,854 (92.4)	337 (78.9)	<0.001	340 (79.6%)	337 (78.9%)	0.866
	1471 (7.6)	90 (21.1)		87 (20.4%)	90 (21.1%)	
stroke	18,356 (95.0)	387 (90.6)	<0.001	385 (90.2%)	387 (90.6%)	0.908
	969 (5.0)	40 (9.4)		42 (9.8%)	40 (9.4%)	
stomach or digestive system disease	17,524 (90.7)	340 (79.6)	<0.001	336 (78.7%)	340 (79.6%)	0.800
	1801 (9.3)	87 (20.4)		91 (21.3%)	87 (20.4%)	
arthritis	17,270 (89.4)	360 (84.3)	0.001	362 (84.8%)	360 (84.3%)	0.925
	2055 (10.6)	67 (15.7)		65 (15.2%)	67 (15.7%)	
memory-related disease	18,906 (97.8)	394 (92.3)	<0.001	395 (92.5%)	394 (92.3%)	1.000
	419 (2.2)	33 (7.7)		32 (7.5%)	33 (7.7%)	
asthma	18,927 (97.9)	405 (94.8)	<0.001	406 (95.1%)	405 (94.8%)	1.000
	398 (2.1)	22 (5.2)		21 (4.9%)	22 (5.2%)	

In KDMI group, 63%, 46.4% and 31.4% of them are subjected to hypertension, hyperlipidemia and diabetes, respectively. 55.3% of them are engaged in vigorous activity. Nearly half are subjected to lower educational backgrounds, 45.2% of them with ≤ primary background. As for the residence, individuals in rural areas account for 72.4%; 82% of them are married. Those who sleep ≤ 5 h constitute 52.7%. More than 50% of them do not smoke (58.1%) or drink (72.6%). The majority of them have no access to the Internet, accounting for a stunning 90.2%. When reporting self-health, few of them are prone to "very good" or "good", making up 2.3% and 3.3%, respectively. Concerning the asset, nearly half of them are less handsome, with 48.5% of them ≤ 3000 Yuan. More detailed descriptions are listed in Table 2.

Table 2. Baseline characteristics of two groups.

Variables	Assignment	Without KDMI (N = 427)	With KDMI (N = 427)
Hypertension	No	155 (36.3%)	158 (37%)
	Yes	272 (63.7%)	269 (63%)
Hyperlipidemia	No	261 (61.1%)	229 (53.6%)
	Yes	166 (38.9%)	198 (46.4%)
Diabetes	No	327 (76.6%)	293 (68.6%)
	Yes	100 (23.4%)	134 (31.4%)
Physical activity	Light	66 (15.5%)	65 (15.2%)
	Moderate	162 (37.9%)	126 (29.5%)
	Vigorous	199 (46.6%)	236 (55.3%)
Sex	Men	199 (46.6%)	190 (44.5%)
	Women	228 (53.4%)	237 (55.5%)
Age	≤ 55 years	64 (15%)	69 (16.2%)
	≤ 65 years	144 (33.7%)	123 (28.8%)
	≤ 75 years	126 (29.5%)	146 (34.2%)
	> 75 years	93 (21.8%)	89 (20.8%)
Education levels	≤ primary	111 (26%)	193 (45.2%)
	≤ high	196 (45.9%)	185 (43.3%)
	≤ college	102 (23.9%)	46 (10.8%)
	> college	18 (4.2%)	3 (0.7%)
Residence	urban	160 (37.5%)	87 (20.4%)
	bounary	37 (8.7%)	29 (6.8%)
	rural	227 (53.2%)	309 (72.4%)
	special area	3 (0.7%)	2 (0.5%)
Martial status	married	351 (82.2%)	350 (82%)
	divorced	16 (3.7%)	8 (1.9%)
	widowed	57 (13.3%)	69 (16.2%)
	never married	3 (0.7%)	0 (0%)
Sleeping duration	≤ 5 h	123 (28.8%)	225 (52.7%)
	≤ 6 h	123 (28.8%)	79 (18.5%)
	≤ 7 h	83 (19.4%)	43 (10.1%)
	≤ 8 h	73 (17.1%)	49 (11.5%)
	> 8 h	25 (5.9%)	31 (7.3%)
Smoking	No	245 (57.4%)	248 (58.1%)
	Yes	182 (42.6%)	179 (41.9%)
Drinking	No	267 (62.5%)	310 (72.6%)
	Yes	160 (37.5%)	117 (27.4%)
Social activity	No	152 (35.6%)	209 (48.9%)
	Yes	275 (64.4%)	218 (51.1%)
Internet access	No	330 (77.3%)	385 (90.2%)
	Yes	97 (22.7%)	42 (9.8%)
BADL	No	332 (77.8%)	272 (63.7%)
	Yes	95 (22.2%)	155 (36.3%)
IADL	No	299 (70%)	226 (52.9%)
	Yes	128 (30%)	201 (47.1%)
Self report	very good	51 (11.9%)	10 (2.3%)
	good	51 (11.9%)	14 (3.3%)
	fair	193 (45.2%)	150 (35.1%)
	poor	93 (21.8%)	179 (41.9%)
	very poor	39 (9.1%)	74 (17.3%)
Asset	≤ 3000 Yuan	124 (29%)	207 (48.5%)
	≤ 8000 Yuan	72 (16.9%)	79 (18.5%)
	> 8000 Yuan	231 (54.1%)	141 (33%)

3.2. Elastic net

In this study, EN was employed for feature selection. The key parameters of optimized model performance were filtered out using the tenfold cross-validation (λ = 0.01147, α = 0.9). The coefficients of the risk factors that are not closely related to KDMI are compressed to 0 and eliminated. The results revealed that 11 factors remained significantly associated with KDMI, including diabetes (0.040), physical activity (0.142), sex (-0.039), age (-0.047), education levels (-0.356), residence (0.109), sleep duration (-0.154), social activity (-0.143), Internet access (-0.022), self-report (0.469) and asset (-0.205). The coefficients were shown in Figure 1. By employing this approach, we focused on the risk factors most closely associated with KDMI, leading to a simplified structure of the Bayesian networks (BNs) model.

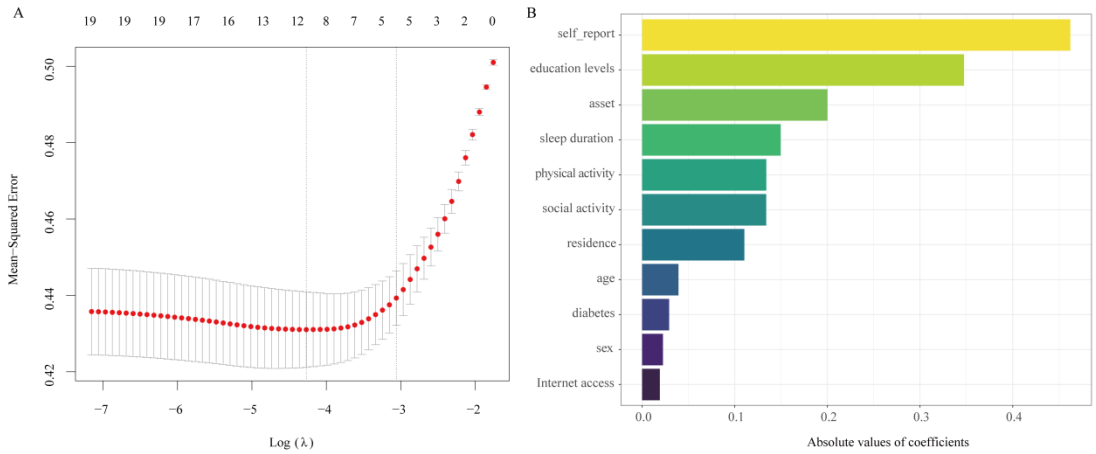

Figure 1. Results of EN and variable coefficient. A: EN showed 11 of the factors associated with KDMI, using the tenfold cross-validation. B: ranking of the coefficients for each variable.

3.3. Bayesian networks

Likewise, the 11 variables and KDMI were included in BNs construction. In this study, BNs were constructed with 12 nodes and 24 directed edges. Nodes represent variables and directed edges represent probabilistic dependence between connected nodes. The percentage in the figure means the prior probability of each node. As shown in Figure 2, the prior probability of KDMI represents 0.50, i.e., P(KDMI) = 0.50. BNs showed that diabetes, physical activity, education levels, sleep duration, social activity, self-report and asset are the parental nodes of KDMI, suggesting that they are directly linked to KDMI. Additionally, sex, age, residence and Internet access represent indirect factors for KDMI. BNs suggests that asset is also indirectly related to KDMI through self-report and sleep duration. It also shows that age could be indirectly related to residence through education levels. Additionally, age appeared to have an indirect impact on KDMI via its association with education levels and physical activity. Notably, Internet access could indirectly relate to KDMI through self-report, social activity and sleep duration. This observation highlights BNs' capability to identify intermediate links between the relevant factors and KDMI.

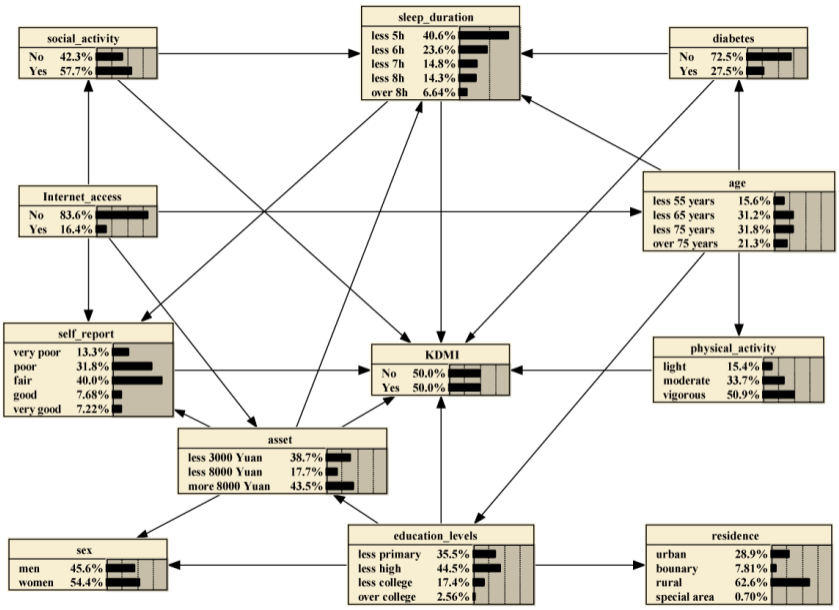

Figure 2. BNs constructed with TS algorithm for KDMI. BNs were constructed with 12 nodes and 24 directed edges. Nodes represent variables, and directed edges represent probabilistic dependence between connected nodes. The percentage in the figure means the prior probability of each node.

3.4. Bayesian reasonings

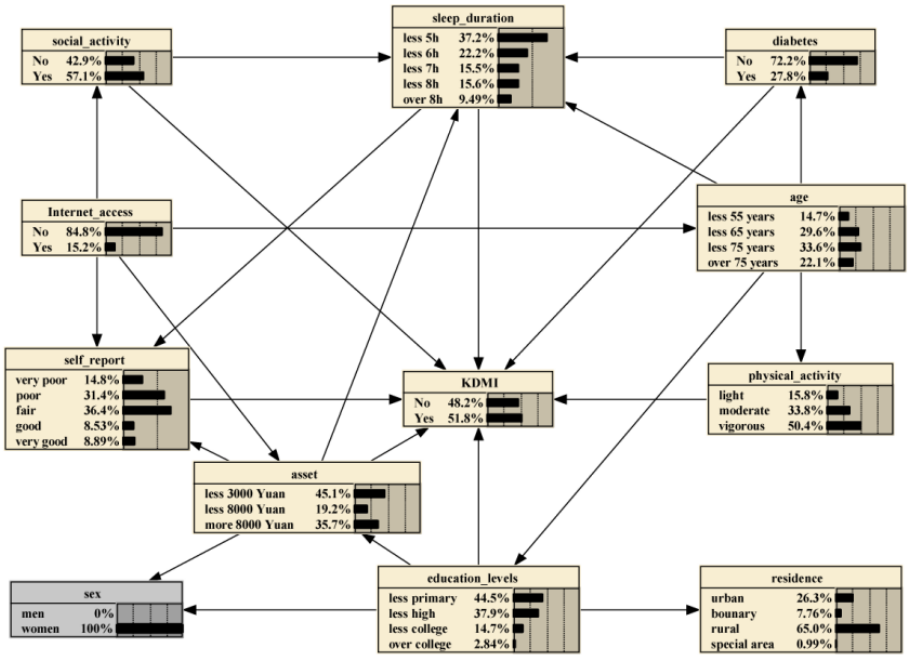

Figure 3. Bayesian reasoning for KDMI for female.

BNs possess the ability to infer unknown nodes based on known nodes, thereby facilitating risk prediction of disease occurrence. Specifically, the probabilistic model allows for quantitative analysis of the influence of various factors on KDMI through the computation of conditional probabilities, denoted as P(y|xi). As shown in Figure 3, for instance, if an individual is female, her probability of developing KDMI increases from the prior probability of 0.50–0.518, i.e., P(KDMI|women) = 0.518. As shown in Figure 4, if the woman has access to the Internet, the probability for her to develop KDMI decreases to 0.488, that is, P (KDMI|women, Internet access) = 0.488. Of note, the probability for her to have social activity, good self-report, asset > 8000 Yuan rises to 0.993, 0.178, 0.562, that is, P(social activity|women, Internet access) = 0.993, P(good self-report|women, Internet access) = 0.178; P(asset > 8000 Yuan|women, Internet access) = 0.562. If the woman is also subjected to sleep duration ≤ 5 hours, moderate physical activity, no social activity, bad self-report on health and less than 3000 Yuan asset, the probability for her to develop KDMI increases to 0.736, that is, P(KDMI|women, Internet access, ≤ 5 h sleep duration, moderate physical activity, no social activity, bad self-report and ≤ 3000 Yuan asset) = 0.736, as shown in Figure 5. Furthermore, the combination of multiple factors such as diabetes, age over 75 years, education background less than primary school, residing in a rural area and no Internet access, as shown in Figure 6, raises her probability of developing KDMI to 0.857, i.e., P(KDMI|women, Internet access, ≤ 5 h sleep duration, moderate physical activity, no social activity, bad self-report, ≤ 3000 Yuan asset, diabetes, age over 75 years, less than primary school, rural area and no Internet access) = 0.857 (Figure 6).

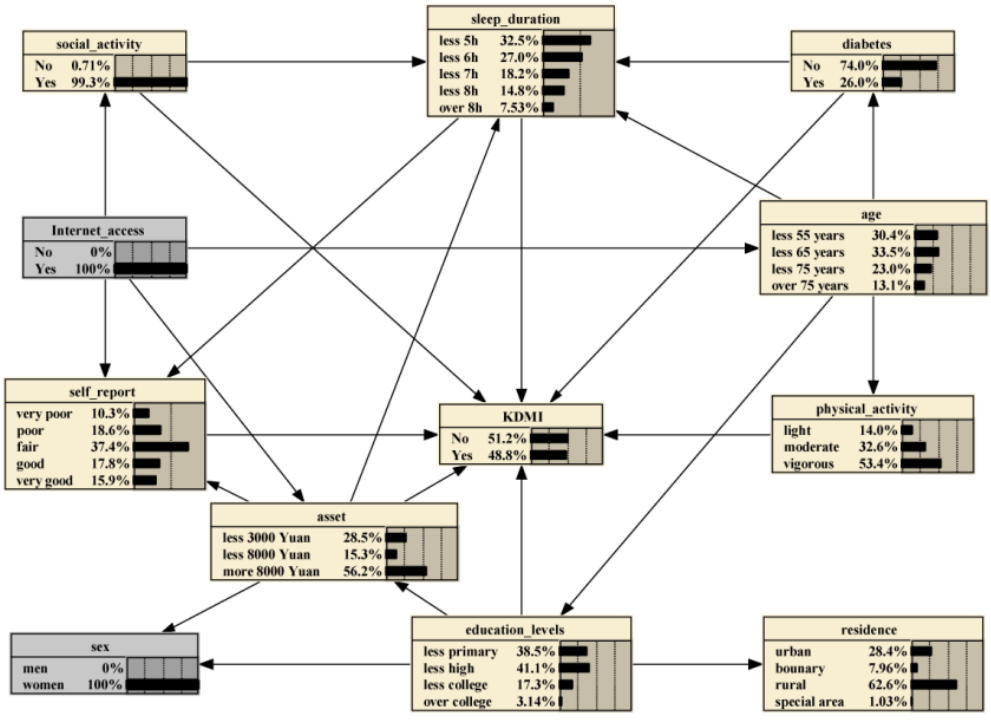

Figure 4. Bayesian reasoning for KDMI under sex of women and Internet access.

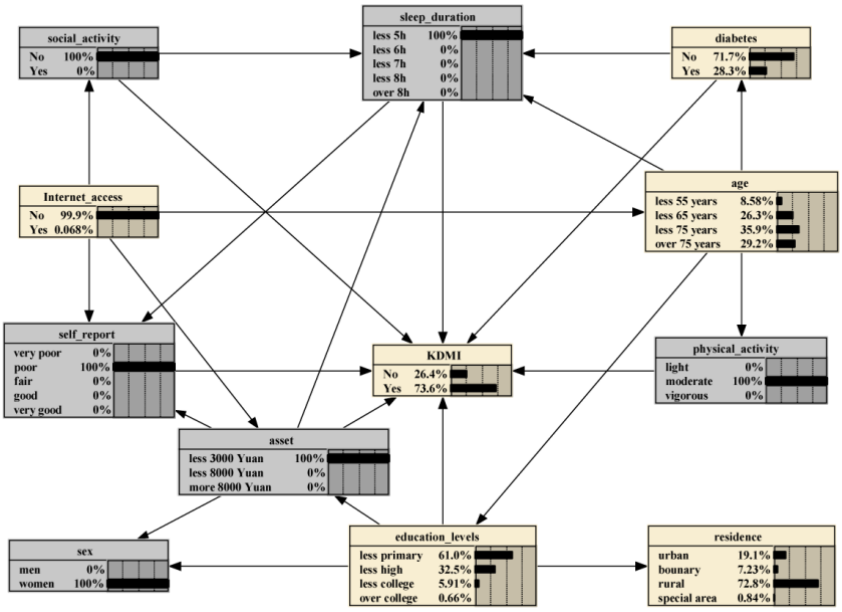

Figure 5. Bayesian reasoning for KDMI under several situations.

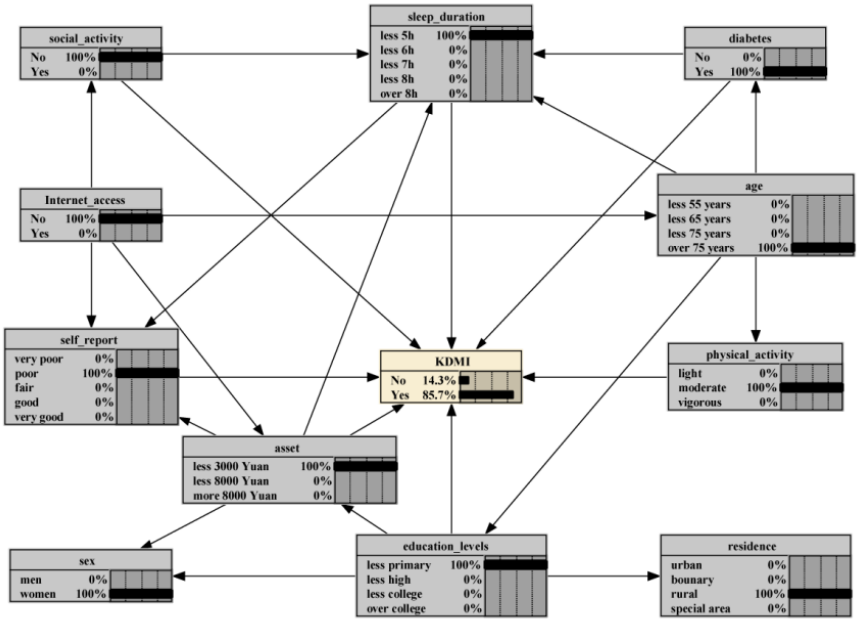

Figure 6. Bayesian reasoning for KDMI under all possible situations.

4. Discussion

In recent years, BNs, a data-driven model, has emerged as an effective approach to discussing factors associated with disease, winning great attention from clinical studies, bioinformatics, decision support systems as well as gaming and law. When utilizing data-driven models for disease prediction, it is crucial to address the issue of unbalanced data before constructing the model, as this can

significantly impact the model's performance. Additionally, considering that a higher number of variables can contribute to a more complex network relationship, careful selection of highly-correlated variables is essential before BNs construction. In this study, we employed BNs with TS algorithm together with PSM and EN to explore factors associated with KDMI. Our results suggest that diabetes, physical activity, education levels, sleep duration, social activity, self-report on health, asset, sex, age, residence and Internet access are highly related to KDMI. Furthermore, we show that the former six ones represent direct factors associated with KDMI, while the remaining five factors exert their influence indirectly on the occurrence of KDMI.

As far as our knowledge is aware, our study is the first one to employ BNs with tabu algorithm to discuss factors associated with KDMI. While scholars typically resort to using LR to investigate KDMI-related factors, LR comes with certain limitations, particularly its inability to perform sequential prediction. In clinical practice, patients may withhold some disease information, making some data unavailable and thus contributing to model dysfunctionality. BNs with Tabu algorithm is superior to traditional LR model.

The first one lies in its capability to maximize data information. BNs with Tabu algorithm offer significant advantages over traditional LR models. First, BNs can maximize data information by accommodating correlated clinical variables, which LR fails to utilize effectively. Being a data-based model with less strict requirements on data distribution, BNs enable comprehensive exploration of collected data, allowing for the identification of both direct and indirect factors associated with the disease. As shown in Table 3, our chi-squared test reveals statistically significant differences between individuals with and without Internet access concerning age, social activity, self-report and asset, suggesting these variables are not inter-dependent and fail to meet LR's prerequisites. In contrast, BNs could show us that Internet access is closely associated with age, social activity, self-report and asset through Bayesian reasoning, flexibly capturing the relationship between variables. As such, BNs could make full use of data information and offer a scientific and insightful idea for disease control and prevention.

The second one concerns the interaction between KDMI and its related factors. Disease occurrence often involves complex interactions between various risk factors. While LR requires additional interaction analysis to assess how risk factors impact the disease, making the analysis more challenging, BNs offer a comprehensive network representation of the intricate relationships between KDMI and its related factors. Additionally, BNs graphically illustrate the interconnections between the related factors, offering insights into their internal relationships. For instance, our results demonstrate that asset indirectly influences KDMI through self-report and sleep duration. Furthermore, Internet access indirectly relates to KDMI through its connections with self-report, social activity and sleep duration, underscoring BNs' ability to identify intermediate links between related factors and KDMI and making it a suitable choice for detecting associated factors in clinical studies. Consequently, the BNs model emerges as a more comprehensive and rational approach in the analysis of factors associated with KDMI in clinical research.

Sleep deprivation can trigger a decrease in kidney function due to mental problems. Diabetic nephropathy, with annually increased prevalence, is primarily responsible for CKD and becomes an important influencing factor for KDMI due to the number of comorbidities after long-term medication. Besides, a national study suggested that severe pruritus, lack of exercise, low income and multiple comorbidities have also been major contributors to emotional problems in maintenance hemodialysis patients [35].

Severe pruritus and multiple complications are associated with CKD, which could be prevented by aggressive treatment. Studies on the mechanisms that lead to depression and anxiety in patients with CKD suggest that they may involve inflammatory states secondary to uremia toxins, oxidative stress due to increased cytokine production and cerebrovascular injury involving the renin-angiotensin system [36]. Therefore, taking certain prevention and treatment measures for various influencing factors can better improve the prognosis of CKD patients. Active control of blood glucose, appropriate physical and social activities, encouragement of increased network communication and improvement of place of residence can improve the mental problems of patients with KDMI, thereby further improving the prognosis of kidney disease. A Japanese study suggests that moderate-intensity exercise of appropriate duration would be an important adjunctive treatment option for patients with CKD or kidney transplantation [37]. Besides, psychological care interventions in hemodialysis patients allow for a significant effect on reducing complication rates, improving anxiety, mental problems, treatment compliance, quality of life and satisfaction with care [38].

This study should be interpreted in the context of several limitations, one of which concerns the BNs model itself. Since the model is data-driven, it could not truly reflect a causal relationship, but a correlation between variables. Second, in this cross-sectional study, some data was collected using questionnaires, which may underestimate the prevalence of CKD and psychiatric diseases. Third, dynamic BNs should be employed for further discussion of the relationship between KDMI and its related factors detected in this study, which will be our ongoing work. Last, this study focused on CKD with psychiatric disease. Our future work should target CKD combined with depression, which is more common in clinical practice.

In short, based on PSM and EN, BNs with tabu algorithm has demonstrated to be an outstanding approach for exploring factors associated with KDMI. The BNs model not only reveals the complex network relationships between KDMI and its associated factors but also facilitates Bayesian reasoning for KDMI, thereby providing valuable insights for guideline development and introducing innovative approaches to clinical practice.

Table 3. Comparisons of some variables between Internet access groups.

Variables	levels	Without Internet access (N = 715)	With Internet access (N = 139)	*P*
age	≤ 55 years	89 (12.4%)	44 (31.7%)	< 0.001
	≤ 65 years	219 (30.6%)	48 (34.5%)	
	≤ 75 years	242 (33.8%)	30 (21.6%)	
	> 75 years	165 (23.1%)	17 (12.2%)	
Social activity	No	361 (50.5%)	0 (0%)	< 0.001
	Yes	354 (49.5%)	139 (100%)	
Self report	very good	44 (6.2%)	17 (12.2%)	< 0.001
	good	40 (5.6%)	25 (18%)	
	fair	276 (38.6%)	67 (48.2%)	
	poor	249 (34.8%)	23 (16.5%)	
	very poor	106 (14.8%)	7 (5%)	
asset	≤ 3000 Yuan	315 (44.1%)	16 (11.5%)	< 0.001
	≤ 8000 Yuan	134 (18.7%)	17 (12.2%)	
	> 8000 Yuan	266 (37.2%)	106 (76.3%)	

5. Conclusions

Risk factor detection plays a crucial role in disease prevention. Our study demonstrates that the implementation of BNs combined with the tabu algorithm yields significant benefits. Notably, it enables the exploration of the intricate network relationship between risk factors and KDMI, while also facilitating risk prediction for KDMI. As a result, our findings offer scientific insights for the control and treatment of KDMI, ultimately contributing to a reduction in its prevalence. The specific findings are outlined below:

1) The BN model for KDMI was constructed, which contains 12 nodes and 24 directed edges. Diabetes, physical activity, education, sleep time, social activity, health self-report and assets are direct factors affecting KDMI, while gender, age, place of residence and internet access are indirect factors affecting KDMI.

2) The BN with tabu algorithm allows probabilistic inference of unknown nodes through known nodes, flexibly demonstrating the impact of a risk factor on KDMI.

3) The BN model with tabu algorithm has great advantages in risk factor detection and has great promise in clinical applications.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

We are grateful to Peking University for providing the free online data. We are also indebted to those who helped us a lot during our writing.

Conflict of interest

The authors declare that they have no competing interests.

References

1. L. Zhang, F. Wang, L. Wang, W. Wang, B. Liu, J. Liu, et al., Prevalence of chronic kidney disease in China: a cross-sectional survey, *Lancet*, **379** (2012), 815–822. https://doi.org/10.1016/S0140-6736(12)60033-6
2. S. Palmer, M. Vecchio, J. C. Craig, M. Tonelli, D. W. Johnson, A. Nicolucci, et al., Prevalence of depression in chronic kidney disease: systematic review and meta-analysis of observational studies, *Kidney Int.*, **84** (2013), 179–191. https://doi.org/10.1038/ki.2013.77
3. F. Teles, V. F. Azevedo, C. T. Miranda, M. P. Miranda, M. C. Teixeira, R. M. Elias, Depression in hemodialysis patients: the role of dialysis shift, *Clinics*, **69** (2014), 198–202. https://doi.org/10.6061/clinics/2014(03)10
4. L. Cirillo, R. Cutruzzulà, C. Somma, M. Gregori, G. Cestone, C. Pizzarelli, et al., Depressive symptoms in dialysis: prevalence and relationship with uremia-related biochemical parameters, *Blood Purif.*, **46** (2018), 286–291. https://doi.org/10.1159/000491014

5. S. C. Palmer, M. Vecchio, J. C. Craig, M. Tonelli, D. W. Johnson, A. Nicolucci, et al., Association between depression and death in people with CKD: a meta-analysis of cohort studies, *Am. J. Kidney Dis.*, **62** (2013), 493–505. https://doi.org/10.1053/j.ajkd.2013.02.369
6. C. Cogley, C. Carswell, K. Bramham, J. Chilcot, Chronic kidney disease and severe mental illness: addressing disparities in access to health care and health outcomes, *Clin. J. Am. Soc. Nephrol.*, **17** (2022), 53–61. https://doi.org/10.2215/CJN.15691221
7. A. J. Kogon, J. Y. Kim, N. Laney, J. Radcliffe, S. R. Hooper, S. L. Furth, et al., Depression and neurocognitive dysfunction in pediatric and young adult chronic kidney disease, *Pediatr. Nephrol.*, **34** (2019), 1575–1582. https://doi.org/10.1007/s00467-019-04265-z
8. D. Duan, L. Yang, M. Zhang, X. Song, W. Ren, Depression and associated factors in Chinese patients with chronic kidney disease without dialysis: a cross-sectional study, *Front. Public Health*, **9** (2021), 605651. https://doi.org/10.3389/fpubh.2021.605651
9. B. M. Melnyk, S. A. Kelly, J. Stephens, K. Dhakal, C. McGovern, S. Tucker, et al., Interventions to improve mental health, well-being, physical health, and lifestyle behaviors in physicians and nurses: a systematic review, *Am. J. Health Promot.*, **34** (2020), 929–941. https://doi.org/10.1177/0890117120920451
10. K. A. McDougall, J. W. Larkin, R. L. Wingard, Y. Jiao, S. Rosen, L. Ma, et al., Depressive affect in incident hemodialysis patients, *Clin. Kidney J.*, **11** (2018), 123–129. https://doi.org/10.1093/ckj/sfx054
11. F. Farrokhi, N. Abedi, J. Beyene, P. Kurdyak, S. V. Jassal, Association between depression and mortality in patients receiving long-term dialysis: a systematic review and meta-analysis, *Am. J. Kidney Dis.*, **63** (2014), 623–635. https://doi.org/10.1053/j.ajkd.2013.08.024
12. A. A. Lopes, J. Bragg, E. Young, D. Goodkin, D. Mapes, C. Combe, et al., Depression as a predictor of mortality and hospitalization among hemodialysis patients in the United States and Europe, *Kidney Int.*, **62** (2002), 199–207. https://doi.org/10.1046/j.1523-1755.2002.00411.x
13. W. Song, H. Gong, Q. Wang, L. Zhang, L. Qiu, X. Hu, et al., Using Bayesian networks with Max-Min Hill-Climbing algorithm to detect factors related to multimorbidity, *Front. Cardiovasc. Med.*, **9** (2022), 984883. https://doi.org/10.3389/fcvm.2022.984883
14. P. Schober, T. R.Vetter, Logistic regression in medical research, *Anesth. Analg.*, **132** (2021), 365–366. https://doi.org/10.1213/ANE.0000000000005247
15. M. Iwagami, K. E. Mansfield, J. F. Hayes, K. Walters, D. P. Osborn, L. Smeeth, et al., Severe mental illness and chronic kidney disease: a cross-sectional study in the United Kingdom, *Clin. Epidemiol.*, **10** (2018), 421–429. https://doi.org/10.2147/CLEP.S154841
16. Z. Zhang, J. Zhang, Z. Wei, H. Ren, W. Song, J. Pan, et al., Application of tabu search-based Bayesian networks in exploring related factors of liver cirrhosis complicated with hepatic encephalopathy and disease identification, *Sci. Rep.*, **9** (2019), 6251. https://doi.org/10.1038/s41598-019-42791-w
17. 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
18. 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

19. W. Song, L. Qiu, J. Qing, W. Zhi, Z. Zha, X. Hu, et al., Using Bayesian network model with MMHC algorithm to detect risk factors for stroke, *Math. Biosci. Eng.*, **19** (2022), 13660–13674. https://doi.org/10.3934/mbe.2022637
20. 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
21. 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
22. F. Castelletti, L. L. Rocca, S. Peluso, F. C. Stingo, G. Consonni, Bayesian learning of multiple directed networks from observational data, *Stat. Med.*, **39** (2020), 4745–4766. https://doi.org/10.1002/sim.8751
23. J. S. Eswari, K. Kavya, Optimal feed profile for the Rhamnolipid kinetic models by using Tabu search: metabolic view point, *AMB Express*, **6** (2016), 116. https://doi.org/10.1186/s13568-016-0279-8
24. M. S. Furqan, M. Y. Siyal, Elastic-net copula granger causality for inference of biological networks, *PLoS One*, **11** (2016), e0165612. https://doi.org/10.1371/journal.pone.0165612
25. W. Song, X. Zhou, Q. Duan, Q. Wang, Y. Li, A. Li, et al., Using random forest algorithm for glomerular and tubular injury diagnosis, *Front. Med.*, **9** (2022), 911737. https://doi.org/10.3389/fmed.2022.911737
26. P. Schober, T. R. Vetter, Propensity score matching in observational research, *Anesth. Analg.*, **130** (2020), 1616–1617. https://doi.org/10.1213/ANE.0000000000004770
27. L. Zhou, X. Ma, W. Wang, Relationship between cognitive performance and depressive symptoms in Chinese older adults: the China health and retirement longitudinal study (CHARLS), *J. Affect Disord.*, **281** (2021), 454–458. https://doi.org/10.1016/j.jad.2020.12.059
28. B. L. Pergola, S. Moonie, J. Pharr, T. Bungum, J. L. Anderson, Sleep duration associated with cardiovascular conditions among adult Nevadans, *Sleep Med.*, **34** (2017), 209–216. https://doi.org/10.1016/j.sleep.2017.03.006.Z
29. V. Sember, K. Meh, M. Sorić, G. Starc, P. Rocha, G. Jurak, Validity and reliability of international physical activity questionnaires for adults across EU countries: systematic review and Meta analysis, *Int. J. Environ. Res. Public Health*, **17** (2020), E7160. https://doi.org/10.3390/ijerph17197161
30. 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
31. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. B*, **67** (2005), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
32. D. Kaur, M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, et al., Application of Bayesian networks to generate synthetic health data, *J. Am. Med. Inf. Assoc.*, **28** (2021), 801–811. https://doi.org/10.1093/jamia/ocaa303
33. F. Glover, Tabu search—Part I, *INFORMS J. Comput.*, **2** (1990), 4–32. https://doi.org/10.1287/ijoc.2.4.4
34. W. Song, Z. Qin, X. Hu, H. Han, A. Li, X. Zhou, et al., Using Bayesian networks with Tabu-search algorithm to explore risk factors for hyperhomocysteinemia, *Sci. Rep.*, **13** (2023), 1610. https://doi.org/10.1038/s41598-023-28123-z

35. Y. Meng, H. T. Wu, J. L. Niu, Y. Zhang, H. Qin, L. L. Huang, et al., Prevalence of depression and anxiety and their predictors among patients undergoing maintenance hemodialysis in Northern China: a cross-sectional study, *Renal Failure*, **44** (2022), 933–944. https://doi.org/10.1080/0886022X.2022.2077761
36. D. S. Kim, S. W. Kim, H. W. Gil., Emotional and cognitive changes in chronic kidney disease, *Korean J. Int. Med.*, **37** (2022), 489–501. https://doi.org/10.3904/kjim.2021.492
37. H. Arazi, M. Mohabbat, P. Saidie, A. Falahati, K. Suzuki., Effects of different types of exercise on kidney diseases, *Sport*, **10** (2022), 42. https://doi.org/10.3390/sports10030042
38. M. Bossola, G. Pepe, M. Antocicco, A. Severino, E. Di Stasio, Interdialytic weight gain and educational/cognitive, counseling/behavioral and psychological/affective interventions in patients on chronic hemodialysis: a systematic review and meta-analysis, *J. Nephrol.*, **35** (2022), 1973–1983. https://doi.org/10.1007/s40620-022-01450-6

AIMS Press