BMC Gastroenterology

RESEARCH

Open Access

Factors influencing the incidence of early gastric cancer: a bayesian network analysis

Ruiyu Li[1], Taiming Yang[2], Zi Dong[2], Yin Gao[1], Nan Li[1], Ting Song[1], Jinshu Sun[2] and Ying Chen[1*]

Abstract

Background This study aims to establish a Bayesian network risk prediction model for gastric cancer using data mining methods. It explores both direct and indirect factors influencing the incidence of gastric cancer and reveals the interrelationships among these factors.

Methods Data were collected from early cancer screenings conducted at the People's Hospital of Lincang between 2022 and 2023. Initial variable selection was performed using Least Absolute Shrinkage and Selection Operator (Lasso) and Sliding Windows Sequential Forward Selection (SWSFS), and the screened variables and demographic characteristics features were used as variables for constructing the Bayesian network (BN) model. Subsequently, the performance of the models was evaluated, and the optimal model was selected for network mapping and Bayesian inference using the best model.

Results The incidence rate of gastric cancer in this region's high-risk population was determined to be 7.09%. The BN model constructed from the set of variables consisting of Lasso's selection variables and demographic characteristics had better performance. A total of 12 variables were incorporated into the BN model to form a network structure consisting of 13 nodes and 18 edges. The model shows that age, gender, ethnicity, current address, upper gastrointestinal symptoms (nausea, acid reflux, vomiting), alcohol consumption, smoking, SGIM gastritis, and family history are important risk factors for gastric cancer development.

Conclusion The Bayesian network model provides an intuitive framework for understanding the direct and indirect factors contributing to the early onset of gastric cancer, elucidating the interrelationships among these factors. Furthermore, the model demonstrates satisfactory predictive performance, which may facilitate the early detection of gastric cancer and enhance the levels of early diagnosis and treatment among high-risk populations.

Keywords Gastric cancer, Bayesian networks, Machine learning, Lasso regression

*Correspondence:
Ying Chen
chenying2128@126.com
[1]Yunnan Provincial Key Laboratory of Public Health and Biosafety & School of Public Health, Kunming Medical University, Kunming 650500, Yunnan, China
[2]Department of Gastroenterology, The People's Hospital of Lincang, Lincang 677000, Yunnan, China

Introduction

Gastric cancer (GC) is a malignant tumor characterized by the abnormal proliferation of epithelial cells in the gastric mucosa, making it one of the most prevalent malignancies worldwide and posing a significant threat to human health [1, 2]. According to the latest report from the International Agency for Research on Cancer (IARC), there were approximately 19.96 million new cancer cases and 9.74 million cancer-related deaths globally in 2022. Among these, the number of new gastric cancer cases was 966,000, with nearly 660,000 deaths, ranking fifth in both incidence and mortality worldwide [2]. In China, there were 358,700 new gastric cancer cases and 260,400 deaths, positioning its incidence as the fifth highest among all cancers and its mortality as the third highest [3]. Although the incidence and mortality rates of gastric cancer in China have exhibited a downward trend from 1990 to 2019, in 2022, the country accounted for 37.0% of the world's new gastric cancer cases and 39.4% of gastric cancer deaths, indicating that the overall burden of incidence and mortality remains substantial [4, 5]. Gastric cancer has emerged as a major public health challenge in China.

The five-year survival rate for patients with early-stage gastric cancer can reach as high as 85% [6]. However, due to the subtle nature of early symptoms, most patients are diagnosed at more advanced stages [7]. Even with prompt treatment at this stage, the five-year survival rate declines to below 10% [6]. Consequently, the implementation of gastric cancer screening is essential for improving prognosis and enhancing patient survival rates [8, 9]. Endoscopic examination combined with pathological tissue biopsy is regarded as the "gold standard" for early gastric cancer screening [10–12]. However, imaging and endoscopic procedures often entail high costs and depend on advanced equipment and physician expertise, rendering them difficult to implement in rural areas and regions with limited medical resources [13]. Furthermore, the discomfort associated with endoscopic procedures leads to low acceptance among asymptomatic patients, impeding widespread application [14]. In China, gastric cancer screening primarily targets symptomatic patients and relies on opportunistic endoscopic screenings. The uneven socioeconomic development and healthcare resource allocation across various regions in China have posed significant challenges to the comprehensive implementation of endoscopic screening within the population [15]. Given the low compliance and accessibility of endoscopic screening, investigating the factors influencing the incidence of early gastric cancer and developing risk models can offer valuable support for early intervention strategies.

The progression of gastric cancer is a multifaceted process influenced by a plethora of risk factors [2, 8]. Currently, a prevalent analytical approach involves the utilization of univariate analyses to identify such influencing factors, followed by the construction of models to elucidate the resultant outcomes [16–18]. While this methodology holds merit, it may fall short in precisely capturing the intricate relationships between influencing factors and outcome events, owing to challenges such as overfitting and covariance. To address these limitations, machine learning methods, notably Least Absolute Shrinkage and Selection Operator (Lasso) [19] and Sliding Windows Sequential Forward Selection (SWSFS) [20], present a robust alternative. These methodologies efficiently and accurately pinpoint the factors that genuinely exert an impact on the outcome variable, while mitigating issues like multicollinearity among variables. Incorporating the selected variables into apt models for predictive analysis and interpretation plays a pivotal role in unraveling the underlying mechanisms of disease occurrence. Bayesian networks (BN) constitute a valuable machine learning tool in the realm of medical science for risk prediction. In contrast to traditional models, BN leverages probabilistic reasoning theory to encapsulate multiple variables and their intricate interrelationships, enabling more refined risk stratification. Furthermore, they illustrate the network structure of variable relationships and facilitate the construction of personalized risk prediction models [21, 22]. Such capabilities underscore the significance of BN in advancing our understanding and management of gastric cancer risk.

Lincang City, situated in the southwestern region of Yunnan Province, is characterized by its diverse population, which includes 23 ethnic minorities. In recent years, there has been a notable increase in both the incidence and mortality rates of malignant tumors in this city. According to tumor monitoring data from 2019, Lincang City reported 3,080 new cases of malignant tumors, resulting in an incidence rate of 181.71 per 100,000 individuals. Furthermore, there were 1,447 deaths attributed to malignant tumors, yielding a mortality rate of 85.37 per 100,000 individuals [23]. Importantly, the incidence rate of gastric cancer in Lincang City was found to be 1.62 times higher than the average incidence rate in Yunnan Province, while the mortality rate was 1.80 times higher than the provincial average [24]. Consequently, addressing the rising incidence and mortality rates of gastrointestinal malignant tumors, particularly gastric cancer, has emerged as a pressing public health challenge that requires immediate attention in Lincang City. This study employs machine learning, including Lasso, SWSFS, and BN models, to assess and predict the risk of gastric cancer, aiming to provide more accurate support for clinical early diagnosis and prevention.

Methods

Study design and settings

We conducted a cross-sectional study with 2022–2023 in the People's Hospital of Lincang to screen for diseases and questionnaires among people at high risk of gastric cancer. The definition of people at high risk of gastric cancer was referred to the Gastric Cancer Screening and Early Diagnosis and Treatment Programme (2024 Edition) (hereinafter referred to as the Programme) issued by the National Health Commission of the PRC [25]. The screening was performed in accordance with the objectives and methodology provided for in the programme. Initially, we included all eligible high-risk groups while excluding those who were younger than 40 years of age, non-residents of Lincang city (< = 2 years of local residence), and serious mental or physical illnesses or unwillingness to cooperate. This study was approved by the Ethics Committee of Kunming Medical University (Approval No. KMMU2024MEC021) prior to commencement, and was conducted in accordance with the Declaration of Helsinki, and informed consent was obtained from all participants during the study.

Study samples

A total of 1820 gastric cancer high-risk individuals were included in this study, and were divided into the gastric cancer group ($n = 129$) and the non-gastric cancer group ($n = 1691$) according to whether they developed gastric cancer or not. Gastroscopy results in the Guidelines for the Diagnosis and Treatment of Gastric Cancer (2022 Edition) [26] were utilized to determine whether the patient had developed gastric cancer.

Measures

We developed a questionnaire for the assessment of people at risk of upper gastrointestinal cancer in accordance with the Programme. The questionnaire included four main domains: general characteristics (gender, age, ethnicity and current address), upper gastrointestinal symptoms (bloating, heartburn, acid reflux, nausea, hiccup, belching, eating discomfort, upper abdominal pain, and vomiting), risk factors for upper gastrointestinal cancers (smoking, alcohol consumption, scalding hot food, overspeed eating, indoor air pollution, toothlessness, pernicious anemia), precancerous diseases or lesions (severe gastric intestinal metaplasia, menetrier disease, gastric stump cancer10 years (benign disease), gastric stump cancer 6 years (gastric cancer), low grade intraepithelial neoplasia, family history of upper gastrointestinal tract cancer, stomach polyp, gastritis, gastric ulcer) and gastroscopy for gastric tumors.

Statistical analysis

All analyses were undertaken using R4.3.3 and its installation package and Netical software. Firstly, the "glment" and "randomForest" packages were used for Lasso regression and SWSFS to screen for variables that have a significant impact on gastric cancer. Lasso [19], based on the least squares method, allows for the shrinkage of less significant variable coefficients to zero while retaining the coefficients of important variables. Lasso regression effectively addresses multicollinearity issues among variables and identifies those of notable significance. SWSFS [20] is based on the Random Forest algorithm, which is used to rank the importance of the variables and then add the variables one by one according to the importance scores in descending order and re-run the Random Forest analysis once for each new variable to find the number of variables corresponding to the smallest average out-of-bag estimation error rate. At this point, the corresponding number of variables is the best set of variables.

Secondly, to develop our model, we split the data into a 70% training dataset using the "caret" package and a 30% test dataset. Subsequently, in the training dataset, we employed the variables screened by lasso regression and SWSFS to construct a Bayesian network (BN) model with the "bnlearn" package, the hill-climbing algorithm (hc) for structural learning, and great likelihood estimation to learn the parameters. BN models are models that simulate the uncertainty of causal relationships during reasoning processes [27, 28]. The model primarily consists of a directed acyclic graph and a collection of conditional probability tables. In this framework, nodes represent variables, with each node corresponding to a specific variable. Directed arcs connecting two nodes indicate direct probabilistic dependencies, while conditional probabilities express the strength of relationships among the nodes. Sub-nodes are those to which directed arcs point, while their parent nodes are those from which the arrows originate. The direction of the arrows qualitatively describes the relationships among the nodes.

After constructing the models, accuracy, sensitivity and specificity, positive and negative predictive values, and area under the receiver operating curve (AUC) were utilized to assess the discriminative ability of the models. Subsequently, the BN model was constructed using the better-performing models and Netica software was applied to draw the topology of the BN model and carry out Bayesian inference.

Results

Characteristics of the participants

As presented in Table 1, a total of 1,820 eligible patients were enrolled in the study, comprising 743 males (40.82%) and 1,077 females (59.18%), with a mean age of 56.22 ± 7.64 years. Among the study population, 129

Table 1 Participant characteristics and univariate analyses. [n (%)]

	Total (*N* = 1820)	**GC patients (*N* = 129)**	**Non-GC patients (*N* = 1691)**	***P*-value**
Gender				
Male	743(4.82)	57(7.67)	686(92.33)	0.476
Female	1077(59.18)	72(6.69)	1005(93.31)	
Age				
>=55	721(39.62)	56(7.77)	665(92.23)	0.412
<55	1099(6.38)	73(6.64)	1026(93.36)	
Ethnicity				
Han ethnic group	836(45.93)	74(8.85)	762(91.15)	0.001
Wa ethnic group	338(18.57)	28(8.28)	310(91.72)	
Dai ethnic group	242(13.3)	7(2.89)	235(97.11)	
Yi ethnic group	161(8.85)	13(8.07)	148(91.93)	
Other ethnic groups	243(13.35)	7(2.88)	236(97.12)	
Smoking				
No	1439(79.7)	103(7.16)	1336(92.84)	0.910
Yes	381(2.93)	26(6.82)	355(93.18)	
Alcohol consumption				
No	1655(9.93)	124(7.49)	1531(92.51)	0.049
Yes	165(9.7)	5(3.03)	160(96.97)	
Current address				
Southwestern Regions	754(41.43)	50(6.63)	704(93.37)	0.018
Eastern Regions	785(43.13)	48(6.11)	737(93.89)	
Northern Regions	281(15.44)	31(11.03)	250(88.97)	
Bloating				
No	1020(56.4)	69(6.76)	951(93.24)	0.607
Yes	800(43.96)	60(7.50)	740(92.5)	
Heartburn				
No	1556(85.49)	103(6.62)	1453(93.38)	0.078
Yes	264(14.51)	26(9.85)	238(90.15)	
Acid reflux				
No	1458(8.11)	93(6.38)	1365(93.62)	0.024
Yes	362(19.89)	36(9.94)	326(90.06)	
Nausea				
No	1594(87.58)	103(6.46)	1491(93.54)	0.009
Yes	226(12.42)	26(11.50)	200(88.50)	
Hiccup				
No	1698(93.3)	119(7.01)	1579(92.99)	0.755
Yes	122(6.7)	10(8.20)	112(91.80)	
Belching				
No	1663(91.37)	112(6.73)	1551(93.27)	0.081
Yes	157(8.63)	17(10.83)	140(89.17)	
Eating discomfort				
No	1566(86.4)	104(6.64)	1462(93.36)	0.087
Yes	254(13.96)	25(9.84)	229(90.16)	
Upper abdominal pain				
No	649(35.66)	53(8.17)	596(91.83)	0.215
Yes	1171(64.34)	76(6.49)	1095(93.51)	
Vomiting				
No	1755(96.43)	119(6.78)	1636(93.22)	0.022*
Yes	65(3.57)	10(15.38)	55(84.62)	
Scalding hot food				
No	1536(84.4)	109(7.10)	1427(92.90)	1.000
Yes	284(15.6)	20(7.04)	264(92.96)	

Table 1 (continued)

	Total (*N* = 1820)	GC patients (*N* = 129)	Non-GC patients (*N* = 1691)	*P*-value
Overspeed eating				
No	1368(75.16)	100(7.31)	1268(92.69)	0.592
Yes	452(24.84)	29(6.42)	423(93.58)	
Indoor air pollution				
No	1816(99.78)	129(7.10)	1687(92.90)	1.000*
Yes	4(0.22)	0(0.00)	4(100.00)	
Toothlessness				
No	916(5.33)	65(7.10)	851(92.90)	1.000
Yes	904(49.67)	64(7.08)	840(92.92)	
Pernicious anemia				
No	1810(99.45)	127(7.02)	1683(92.98)	0.155*
Yes	10(0.55)	2(20.00)	8(80.00)	
SGIM				
No	1815(99.73)	127(7.00)	1688(93.00)	0.043*
Yes	5(0.27)	2(40.00)	3(60.00)	
MD				
No	1818(99.89)	128(7.04)	1690(92.96)	0.137*
Yes	2(0.11)	1(50.00)	1(50.00)	
GSC10 (benign disease)				
No	1818(99.89)	128(7.04)	1690(92.96)	0.137*
Yes	2(0.11)	1(50.00)	1(50.00)	
GSC6 (gastric cancer)				
No	1813(99.62)	127(7.00)	1686(93.00)	0.083*
Yes	7(0.38)	2(28.57)	5(71.43)	
LGIN				
No	1818(99.89)	129(7.10)	1689(92.90)	1.000*
Yes	2(0.11)	0(0.00)	2(100.00)	
Family history				
No	1798(98.79)	121(6.73)	1677(93.27)	< 0.001*
Yes	22(1.21)	8(36.36)	14(63.64)	
Cardia polyp				
No	1797(98.74)	128(7.12)	1669(92.88)	1.000*
Yes	23(1.26)	1(4.35)	22(95.65)	
Stomach polyp				
No	1305(71.7)	91(6.97)	1214(93.03)	0.840
Yes	515(28.3)	38(7.38)	477(92.62)	
Gastritis				
No	26(1.43)	7(26.92)	19(73.08)	0.002*
Yes	1794(98.57)	122(6.80)	1672(93.20)	
Gastric ulcer				
No	1750(96.15)	123(7.03)	1627(92.97)	0.631
Yes	70(3.85)	6(8.57)	64(91.43)	

Note: **SGIM**: severe gastric intestinal metaplasia; **MD**: menetrier disease; **GSC10**: gastric stump cancer 10 years (benign disease); **GSC6**: gastric stump cancer 6 years (gastric cancer); **LGIN**: low-grade intraepithelial neoplasia; **Family History**: family history of upper gastrointestinal tract cancer. **Smoking**: consumption of one or more cigarettes per day for a duration of six consecutive or cumulative months; **Alcohol Consumption**: consumption exceeding 1 cup per day for women or 2 cups per day for men; **Scalding Hot Food**: Food with a temperature of ≥ 65 ℃; **Overspeed Eating**: completion of a meal in less than 15 min; **Indoor air pollution**: As per GB/T 18,883 – 2022 (Indoor Air Quality Standards); **Southwestern Regions**: including Zhenkang, Gengma, and Cangyuan counties; **Eastern Regions**: including Shuangjiang, Linxiang, and Yunxian counties; **Northern Regions**: including Fengqing and Yongde counties; **Toothlessness**: presence of ≥ 2 tooth losses

Statistical Analysis: Design-based χ^2 (Chi-squared) test was employed. *: Design-based Fisher's exact test was utilized. A p-value ≤ 0.05 was considered statistically significant.

Table 2 Influencing factors of early gastric cancer selected by Lasso regression

Variables	coefficient
Nausea (Yes)	0.013244766
Acid reflux (Yes)	0.00239885
Current address (Northern)	0.011692466
Ethnicity (Dai)	-0.014283516
Ethnicity (Other)	-0.019095206
Vomiting (Yes)	0.015280802
Gastritis (Yes)	-0.101798245
Alcohol consumption (Yes)	-0.001180737
Family history (Yes)	0.189330094
SGIM (Yes)	0.033112917

Note: SGIM: Severe gastric intestinal metaplasia; Family history: Family history of upper gastrointestinal tract cancer; Northern: including Fengqing and Yongde counties

cases of gastric cancer were identified, accounting for 7.09% of the total, while the remaining 1,691 patients (92.91%) were diagnosed with non-gastric cancers. Specifically, the incidence of gastric cancer among males was 57 cases (7.67%), and among females, it was 72 cases (6.69%). Statistical analysis revealed no significant difference in the detection rate of gastric cancer between genders (χ^2 = 0.508, P = 0.476). Notably, statistically significant differences in the detection rate of gastric cancer were observed across different ethnic groups and current addresses. The Han, Wa, and Yi ethnic groups exhibited similar detection rates exceeding 8.00%, which were higher compared to other ethnic groups (χ^2 = 17.921, P = 0.001). Furthermore, higher rates of gastric cancer were observed in Fengqing and Yongde counties, situated in the northern region of Lincang. In terms of lifestyle factors, individuals who consumed alcohol demonstrated a higher likelihood of developing gastric cancer compared to those who abstained from alcohol consumption (χ^2 = 3.884, P = 0.049). Additionally, gastric cancer was more frequently detected in patients presenting with upper gastrointestinal symptoms such as acid reflux (9.94%), nausea (11.50%), vomiting (15.38%), or severe gastrointestinal symptoms (SGIM) (40.00%). Notably, a history of non-gastritis (26.92%) and a family history of upper gastrointestinal tract cancer (36.36%) were also associated with an increased susceptibility to gastric cancer detection.

Screening of risk factors for gastric cancer

The results obtained from variable screening utilizing both Lasso regression and SWSFS are presented as follows: Initially, Lasso regression identified nine clinically pertinent indicators: ethnicity, current address, upper gastrointestinal symptoms encompassing nausea, acid reflux, and vomiting, alcohol consumption, SGIM, gastritis, and family history. These variables exhibited coefficients with positive values denoting risk factors and negative values indicating protective factors, as delineated in Table 2. The graphical representations of the Lasso fitting results are provided in Figs. 1 and 2 of the supplementary materials.

Furthermore, the outcomes of the SWSFS variable screening are depicted in Fig. 3 of the supplementary materials, alongside Fig. 1 in the main text. Specifically, Fig. 3 of the supplementary materials illustrates that

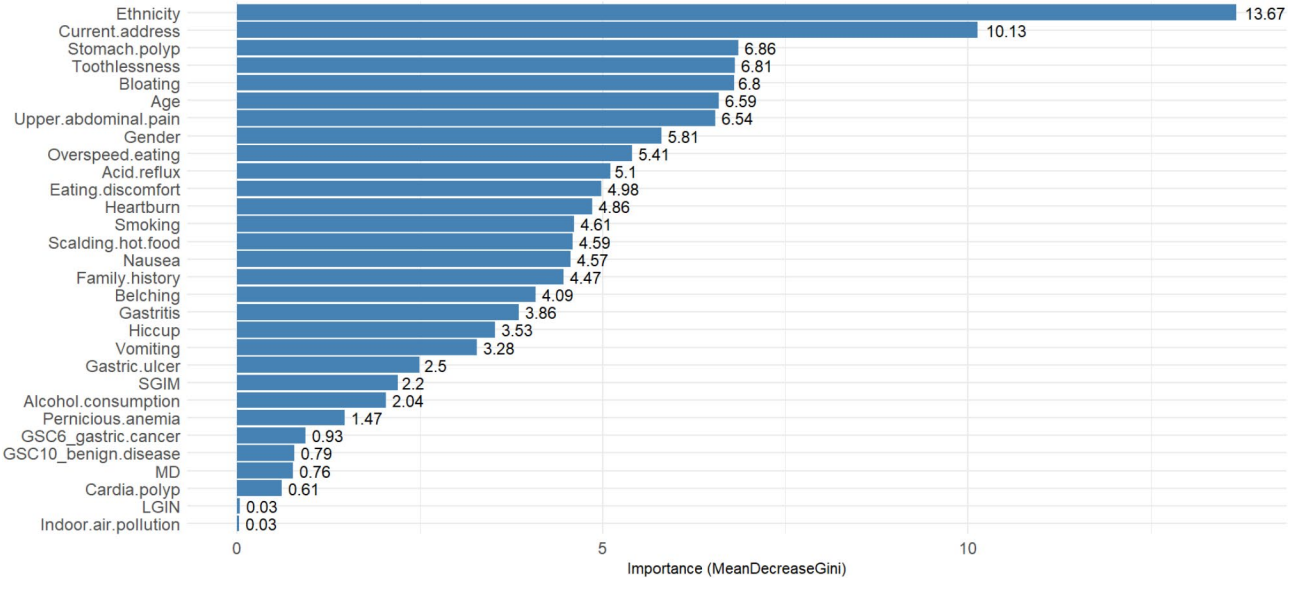

Fig. 1 Importance ranking of influencing factors of early gastric cancer. The horizontal axis is the importance score of the variable, and the vertical axis is the variable name. SGIM: severe gastric intestinal metaplasia; MD: menetrier disease; GSC10: gastric stump cancer 10 years (benign disease); GSC6: gastric stump cancer 6 years (gastric cancer); LGIN: low grade intraepithelial neoplasia; Family history: family history of upper gastrointestinal tract cancer

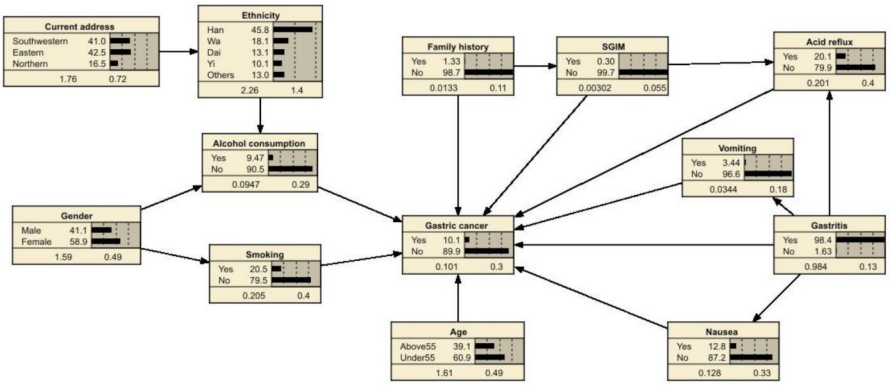

Fig. 2 BN model with 13 nodes and 18 directed edges. Nodes denote variables and directed edges represent probabilistic dependencies between connected nodes. The percentages in the graph represent the prior probability of each node. SGIM: severe gastric intestinal metaplasia; Family history: family history of upper gastrointestinal tract cancer

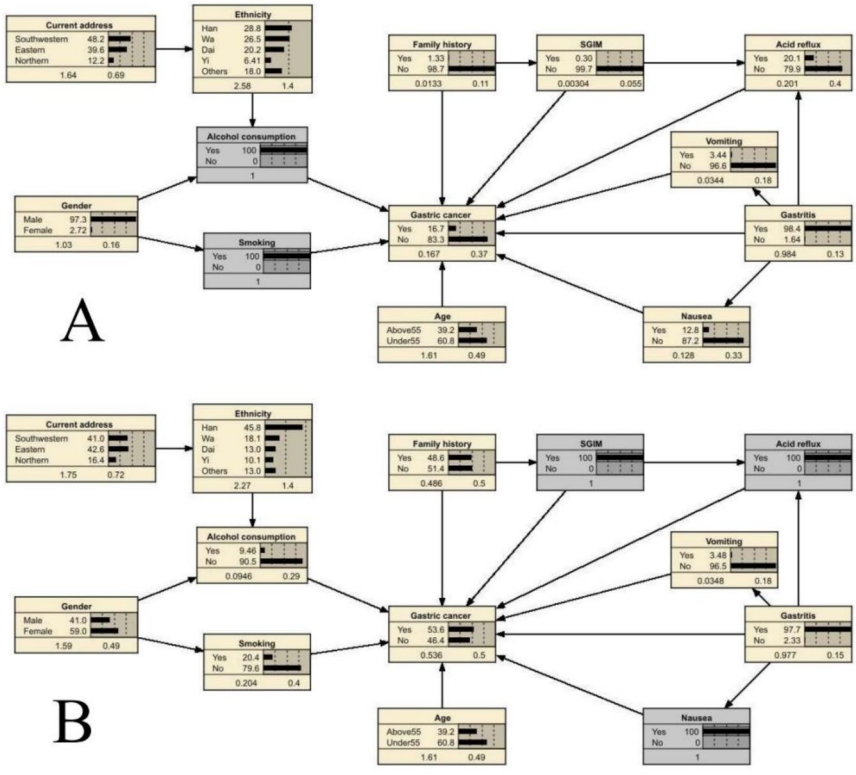

Fig. 3 BN model risk inference for early gastric cancer. **A**. Bayesian inference for subjects who are smoking and drinking alcohol. **B**. Bayesian inference was performed on patients with SGIM and those with chronic upper gastrointestinal symptoms such as acid reflux and nausea. SGIM: severe gastric intestinal metaplasia; Family history: family history of upper gastrointestinal tract cancer

when the number of variables is set to 10, the minimum out-of-bag error achieves a value of 0.0707. According to the ranking based on the importance scores presented in Fig. 1, the top ten variables, in order of significance, are ethnicity, current address, stomach polyp, toothlessness, bloating, age, upper abdominal pain, gender, overspeed eating, and acid reflux.

Construction and evaluation of BN model for early gastric cancer

To gain a deeper understanding of the causal relationships between variables associated with early gastric cancer occurrence, we integrated demographic characteristics of the study subjects, including age, gender, current address, smoking status, and alcohol consumption

Table 3 Overview of the performance estimates for each prediction model

Model	Accuracy	Specificity	Sensitivity	PPV	NPV	AUC
Lasso-BN	0.916	0.974	0.132	0.278	0.937	0.639
SWSFS-BN	0.850	0.892	0.184	0.117	0.936	0.570

Note: PPV: positive predictive value; NPV: negative predictive value; AUC: area under the receiver operating curve

(although not all these variables were identified by Lasso and SWSFS), into the construction of the BN model. The performance metrics of BN models, formulated using variables identified by both methods along with demographic characteristics, are detailed in Table 3. Specifically, the model constructed using variables selected by the Lasso method combined with demographic characteristics accurately represented the relationships between nodes, achieving an AUC of 0.637, an accuracy of 0.916, a specificity of 0.974, a sensitivity of 0.132, a positive predictive value of 0.278, and a negative predictive value of 0.937. To achieve a more precise and comprehensive understanding of the factors influencing gastric cancer occurrence, we constructed a BN model using variables identified by the Lasso method and provided interpretations and reasoning regarding the relationships among these factors.

The BN model constructed using the hc algorithm is depicted in Fig. 2. The resultant network model consists of 13 nodes and 18 edges. Analysis indicates that alcohol consumption, smoking, age, family history, SGIM, vomiting, nausea, acid reflux, and gastritis are directly associated with the occurrence of gastric cancer. Meanwhile, current address, gender, and ethnicity are indirectly related to gastric cancer through their influence on alcohol consumption and smoking. Additionally, family history can indirectly affect the development of gastric cancer via SGIM, while SGIM and gastritis can indicate the presence of gastric cancer through the manifestation of corresponding symptoms.

Inference for the BN model of early gastric cancer

The constructed network model was employed for predictive assessments, utilizing relevant node information to evaluate the risk of gastric cancer in patients. For instance, when a patient has a history of alcohol consumption and smoking but exhibits no other disease symptoms, regardless of ethnicity, BN model inference indicates that the likelihood of developing gastric cancer increases to 16.7% (Fig. 3A). Conversely, when a patient is diagnosed with SGIM and presents with symptoms of nausea and acid reflux, the risk of developing gastric cancer escalates to 53.6% (Fig. 3B).

Discussion

With the aging population and the rising prevalence of unhealthy lifestyles, the burden of gastric cancer in China is gradually increasing [2]. The occurrence and progression of gastric cancer are influenced by multiple factors; identifying these direct and indirect influencing factors can facilitate the early identification of high-risk populations, improve the levels of early diagnosis and treatment [1], and mitigate the loss of quality of life associated with gastric cancer. Consequently, employing a BN model to construct a diagram of the influencing factors elucidates the complex relationships among them, thereby contributing positively to the reduction of gastric cancer incidence.

The findings of the present study reveal that several factors significantly influence the occurrence of gastric cancer, including ethnicity, current address, upper gastrointestinal symptoms (encompassing nausea, acid reflux, and vomiting), alcohol consumption, severity of gastrointestinal symptoms indexed by the SGIM, gastritis, and family history. Notably, age, ethnicity, and current address exhibit a direct correlation with gastric cancer incidence. Consistent with established knowledge, the risk of gastric cancer progressively escalates with advancing age. In the context of Lincang, a city situated on the southwestern border of China, a remarkable concentration of ethnic minorities is observed. Notably, the Wa, Dai, and Yi ethnic groups constitute the primary minorities in this region. These ethnic groups harbor unique customs and dietary habits, such as the consumption of raw meat and untreated water, which may contribute to the heightened incidence of gastric cancer [29]. Furthermore, our analysis indicates that the risk of gastric cancer is more elevated in the northern part of Lincang compared to other regions. This heightened risk may be attributed to the proximity of the northern region to Baoshan and Dehong, areas characterized by high ethnic concentrations and distinctive drinking cultures, ultimately leading to a greater susceptibility to gastric cancer. Consequently, the findings of this study underscore the importance of ethnic areas in the prevention and treatment strategies for gastric cancer.

Consistent with the majority of existing studies, the present investigation proposes that both alcohol consumption and smoking can elevate the risk of gastric cancer, irrespective of gender [8, 30]. This correlation may stem from the role alcohol plays as a solvent in the gastric environment, enabling carcinogenic substances to permeate gastric cells and disrupting the synthesis of prostaglandins and the metabolism of retinoids. Furthermore, alcohol has been implicated in augmenting the generation of bimolecularly toxic free radicals. Notably,

acetaldehyde, a metabolic derivative of alcohol, has been designated as a Group 1 human carcinogen by the International Agency for Research on Cancer (IARC) [31, 32]. However, a meta-analysis of cohort studies examining the relationship between alcohol consumption and gastric cancer risk indicated that light to moderate drinking does not significantly affect the risk of gastric cancer compared to non-drinkers [33]. This phenomenon may be attributed to the antibacterial effects of alcohol against Helicobacter pylori [34, 35]. Furthermore, some research has suggested that male light to moderate drinkers (1–5 g/day) exhibit the lowest risk of alcohol-related cancer mortality [36], implying that moderate alcohol consumption may have a protective effect against gastric cancer. Future studies should aim to further investigate the dose-response relationship between alcohol consumption and the occurrence of gastric cancer, as well as explore the underlying mechanisms involved.

Smoking has been widely recognized as an independent risk factor for the development of gastric cancer [37]. The progression of gastric cancer exhibits a significant dose-response relationship with both the quantity of cigarettes smoked and the duration of smoking. Moreover, the risk of gastric cancer decreases linearly over time since smoking cessation, reaching the level of non-smokers approximately 30 years after quitting [38]. Therefore, extensive publicity and educational campaigns should be carried out to encourage smokers to quit, aiming to reduce the incidence of this disease.

In addition, a positive correlation between smoking and intestinal epithelial hyperplasia has also been observed [38]. SGIM has been identified as a precancerous lesion for gastric cancer [39]. Over an extended period, chronic atrophic gastritis can cause the replacement of gastric mucosal epithelial cells with intestinal epithelial cells, thereby facilitating the transition of precancerous lesions to the intestinal metaplasia stage. Early diagnosis of intestinal metaplasia, followed by appropriate medical interventions and endoscopic treatments, can alleviate clinical symptoms, enhance the physiological function of the gastric mucosa, and reduce or delay the onset of gastric cancer [40].

Interestingly, our study revealed that the detection rate of gastric cancer in patients with gastritis was lower than that in patients without gastritis. This might be attributed to the fact that patients with gastritis may be more vigilant about their lifestyles, taking steps to minimize the occurrence of disease-associated risk factors, which in turn reduces or slows down the development of gastric cancer [41]. This finding suggests that early diagnosis of gastric diseases may enable patients to adopt preventive measures to halt the further progression of the disease. However, it should be noted that this study did not categorize the types of gastritis. Thus, the results should be interpreted with caution, and future research should be more refined to provide a more comprehensive understanding.

Last but not least, acid reflux, vomiting, and nausea were common upper gastrointestinal symptoms associated with various gastrointestinal diseases and served as important predictive indicators for gastric cancer risk. These clinical manifestations may suggest a history of gastrointestinal disorders, such as intestinal metaplasia. A meta-analysis has indicated that the risk of gastric cancer among patients with gastrointestinal diseases and chronic gastrointestinal conditions is 4.85 times and 4.40 times greater, respectively, than that of the general population [42]. Therefore, when patients present with frequent upper gastrointestinal symptoms, further investigation, including gastroscopy and pathological biopsy, when necessary, should be conducted. Furthermore, in alignment with the findings of Zhang Linglin [43] and EOM [44], a family history of upper gastrointestinal cancer is a well-established risk factor for gastric cancer. Studies indicate that 10-15% of gastric cancer patients have a familial history of gastric tumors [45], with a significantly elevated incidence observed among first-degree relatives of gastric cancer patients [46]. This increased risk may be attributed to shared genetic loci within families [47].

Limitations

Despite the contributions of our study, several limitations must be acknowledged. Firstly, the nature of this investigation is cross-sectional, inherently limiting the ability to draw definitive conclusions regarding causality. Nonetheless, our study offers novel insights into the risk factors associated with gastric cancer. Secondly, the sampling frame was restricted to a single institution, potentially limiting the generalizability of our findings to a broader population. To enhance the external validity of future research, it is advisable to involve multiple organizations spanning diverse geographical regions, thereby augmenting the sample size and enabling the collection of a more comprehensive array of relevant indicators. Lastly, our study employed a single model—specifically, the Bayesian Network (BN) model—to elucidate the relationships among the study factors and between these factors and the outcomes. However, predictive performance of the models has to be further improved, and to be able to promote a more nuanced understanding, we provide the dataset used in this study for comparative analyses and interpretations by interested researchers using multiple models.

Conclusion

In this study, Lasso and SWSFS were employed for variable selection, which effectively solved the problem of multicollinearity among variables, thus identifying variables of significance. Subsequently, the BN model was utilized for structural and parameter learning to develop an optimal BN model, allowing for the calculation of the impact of various factors on the early risk of gastric cancer. This approach visually illustrates the direct and indirect relationships among the influencing factors, clarifying the internal regulatory mechanisms that govern these relationships. By overcoming the limitations of traditional predictive models in explaining causal relationships and probability calculations, this methodology aids healthcare professionals in the early identification of high-risk populations for gastric cancer. Furthermore, it enhances the levels of early diagnosis and treatment, facilitating timely targeted interventions to mitigate the risk of gastric cancer and reduce the associated loss of quality of life.

Abbreviations

GC	Gastric cancer
Lasso	Least Absolute Shrinkage and Selection Operator
BN	Bayesian Networks
SGIM	Severe gastric intestinal metaplasia
MD	Menetrier disease
GSC10	Gastric stump cancer 10 years (benign disease)
GSC6	Gastric stump cancer 6 years (gastric cancer)
LGIN	Low grade intraepithelial neoplasia
PPV	Positive predictive value
NPV	Negative predictive value
AUC	Area under the receiver operating curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12876-025-03765-7.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

We gratefully acknowledge the collaborative efforts of the study participants and data collectors in the designated survey area.

Author contributions

RYL, TMY, and YC conceived the study and designed the protocol. ZD, YG, and NL supervised the implementation of the study and advised on data analysis. RYL, TMY, TS, and JSS were involved in the development of the questionnaires, data collection and analysis. RYL drafted the original manuscript. YC revised the original manuscript. All authors reviewed and approved the final manuscript. YC and RD are the sponsors of this study.

Funding

Role of funding source This study was supported by a grant from Yunnan Provincial Talent Program for Young Scholar and Technical Reserve Personnel (202305AC160046), Key Scientific and Technological Project for Sustainable Development Demonstration Zones (202104AC100001-A11), First-Class Discipline Team of Kunming Medical University National (2024XKTDTS16).

Data availability

Availability of data and materialsThe dataset used and/or analyzed in this study can be found in the data of the supplementary material.

Declarations

Ethics approval and consent to participate

The study was performed in strict accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Kunming Medical University (approval number: KMMU2024MEC021). Adhering to the principle of voluntary participation, all potential participants were given the opportunity to make an informed decision on their participation in the study. All subjects were clearly informed of their right to withdraw from the study at any time without facing adverse consequences. To ensure transparency, the purpose and procedures of the study were fully explained to the subjects before signing the informed consents. Individually identifiable information, such as name and telephone number, was deliberately omitted from the recorded data during the data collection phase to ensure anonymity. Finally, Information collected was subjected to appropriate coding procedures and kept strictly confidential throughout the research process.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 27 October 2024 / Accepted: 6 March 2025
Published online: 21 March 2025

References

1. Machlowska J, Baj J, Sitarz M, et al. Gastric cancer: epidemiology, risk factors, classification, genomic characteristics and treatment strategies [J]. Int J Mol Sci. 2020;21(11). https://doi.org/10.3390/ijms21114012.
2. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. CA Cancer J Clin. 2024;74(3):229–63. https://doi.org/10.3322/caac.21834.
3. Zheng RS, Chen R, HaN BF, et al. Cancer incidence and mortality in China, 2022 [J]. Zhonghua Zhong Liu Za Zhi. 2024;46(3):221–31. https://doi.org/10.3760/cma.j.cn112152-20240119-00035.
4. Yifei Y, Kexin S, Rongshou Z. Interpretation and analysis of the Global Cancer Statistics Report 2022: a comparison between China and the world [J]. Chin J Bases Clin Gen Surg. 2024;31(07):769–80. https://doi.org/10.7507/1007-9424.202406046.
5. Zhuoyu L, Weihan K. Epidemiologic features and trends of gastric cancer in the world and China: interpretation of the GLOBOCAN 2018–2022 [J]. Chin J Bases Clin Gen Surg. 2024;1–10. https://doi.org/10.7507/1007-9424.202409074.
6. Yang L, Ying X, Liu S, et al. Gastric cancer: epidemiology, risk factors and prevention strategies [J]. Chin J Cancer Res. 2020;32(6):695–704. https://doi.org/10.21147/j.issn.1000-9604.2020.06.03.
7. Cristescu R, Lee J, Nebozhyn M, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes [J]. Nat Med. 2015;21(5):449–56. https://doi.org/10.1038/nm.3850.
8. Association C G C A O C A-C, Association C S O U G S O C M D, Group C H R M C-G. C. Chinese guideline on risk management of gastric Cancer in the general Public(2023 Edition) [J]. Chin Med J. 2023;103(36):2837–49. https://doi.org/10.3760/cma.j.cn112137-20230608-00968.
9. Jin G, Lv J, Yang M, et al. Genetic risk, incident gastric cancer, and healthy lifestyle: a meta-analysis of genome-wide association studies and prospective cohort study [J]. Lancet Oncol. 2020;21(10):1378–86. https://doi.org/10.1016/S1470-2045(20)30460-5.
10. Sugano K. Screening of gastric cancer in Asia [J]. Best Pract Res Clin Gastroenterol. 2015;29(6):895–905. https://doi.org/10.1016/j.bpg.2015.09.013.
11. Ma E, Sasazuki S, Shimazu T, et al. Reactive oxygen species and gastric cancer risk: a large nested case-control study in Japan [J]. Eur J Epidemiol. 2015;30(7):589–94. https://doi.org/10.1007/s10654-015-0025-6.

12. Bourke MJ, Bergman N. Endoscopic submucosal dissection: indications and application in Western endoscopy practice [J]. Gastroenterology. 2018;154(7):1887–900. https://doi.org/10.1053/j.gastro.2018.01.068.
13. Zhang X, Li M, Chen S, et al. Endoscopic screening in Asian countries is associated with reduced gastric Cancer mortality: A Meta-analysis and systematic review [J]. Gastroenterology. 2018;155(2):347–54. https://doi.org/10.1053/j.gastro.2018.04.026.
14. Fan X, Qin X, Zhang Y, et al. Screening for gastric cancer in China: advances, challenges and visions [J]. Chin J Cancer Res. 2021;33(2):168–80. https://doi.org/10.21147/j.issn.1000-9604.2021.02.05.
15. Zhongxia C, Lin J. Current status and thinking of screening of early gastric cancer [J]. Chin J Dig Endoscopy. 2019;36(5):305–9. https://doi.org/10.3760/cma.j.issn.1007-5232.2019.05.001.
16. Zhu X, Ge B. Analysis of multiple factors influencing the survival of patients with advanced gastric cancer [J]. Aging. 2024;16(10):8541–51. https://doi.org/10.18632/aging.205820.
17. Zhang R, Li H, Li N, et al. Risk factors for gastric cancer: a large-scale, population-based case-control study [J]. Chin Med J (Engl). 2021;134(16):1952–8. https://doi.org/10.1097/CM9.0000000000001652.
18. Yan B B, Cheng L N, Yang H, et al. Comprehensive analysis of risk factors associated with submucosal invasion in patients with early-stage gastric cancer [J]. World J Gastroenterol. 2024;30(47):5007–17. https://doi.org/10.3748/wjg.v30.i47.5007.
19. Goeman JJ. L1 penalized Estimation in the Cox proportional hazards model [J]. Biom J. 2010;52(1):70–84. https://doi.org/10.1002/bimj.200900028.
20. Jiang R, Tang W, Wu X, et al. A random forest approach to the detection of epistatic interactions in case-control studies [J]. BMC Bioinformatics. 2009;10(Suppl 1). https://doi.org/10.1186/1471-2105-10-S1-S65. S65.
21. Wang M, Wang C J, Gu H Q, et al. Sex differences in Short-Term and Long-Term outcomes among patients with acute ischemic stroke in China [J]. Stroke. 2022;53(7):2268–75. https://doi.org/10.1161/STROKEAHA.121.037121.
22. Xuchun W, Weimei S, Mengmeng Z, et al. Related factors of liver cirrhosis complicated with hepatic encephalopathy based on elastic net and bayesian network model [J]. Mod Prev Med. 2021;48(09):1705–9.
23. Menglin F. Study on the prevalence of malignant tumors and disease burden in Linxiang district from 2015–2019 [D]. Huazhong University of Science and Technology; 2022.
24. Qiang Z, Yunchao H. Analysis of Cancer risk assessment and screening results among urban residents in Kunming City [J]. China Cancer. 2018;27(09):641–6.
25. PRC N H C, O T. Gastric Cancer Screening and Early Diagnosis and Treatment Programme. (2024 Edition) [J]. Chinese Journal of Oncology, 2024, 46(10): 915-6.https://doi.org/10.3760/cma.j.cn112152-20240705-00276
26. Bureau of Medical Administration N H C O T P S R O, C. Standardization for diagnosis and treatment of gastric cancer (2022 edition) [J]. Chin J Dig Surg. 2022;21(09):1137–64. https://doi.org/10.3760/cma.j.cn115610-20220726-00432.
27. Topuz K, Davazdahemami B. A bayesian belief network-based analytics methodology for early-stage risk detection of novel diseases [J]. Ann Oper Res. 2023;1–25. https://doi.org/10.1007/s10479-023-05377-4.
28. Meiyan Y, Yaning Z. Influencing factors of H-type hypertension in middle-aged and elderly based on bayesian network model [J]. Mod Prev Med. 2024;51(02):335–42. https://doi.org/10.20043/j.cnki.MPM.202308236.
29. Likun L, Youguo D, Hongmei S, et al. Clinical epidemiological analysis of 1524 patients with gastric cancer in Yunnan tumor hospital from 2016 to 2021 [J]. Volume 44. Medicine and Pharmacy of Yunnan; 2023. pp. 2–5. 01.
30. Jie H, Wanqing C. China guideline for the screening, early detection and early treatment of gastric cancer (2022, Beijing) [J]. China Cancer. 2022;31(07):488–527.
31. Boffetta P. Alcohol and cancer [J]. Lancet Oncol. 2006;7(2):149–56. https://doi.org/10.1016/S1470-2045(06)70577-0.
32. Bouras E, Tsilidis K K, Triggi M, et al. Diet and risk of gastric Cancer. Umbrella Rev [J] Nutrients. 2022;14(9). https://doi.org/10.3390/nu14091764.
33. He Z, Zhao T T, Xu H M, et al. Association between alcohol consumption and the risk of gastric cancer: a meta-analysis of prospective cohort studies [J]. Oncotarget. 2017;8(48):84459–72. https://doi.org/10.18632/oncotarget.20880.
34. Ogihara A, Kikuchi S, Hasegawa A, et al. Relationship between Helicobacter pylori infection and smoking and drinking habits [J]. J Gastroenterol Hepatol. 2000;15(3):271–6. https://doi.org/10.1046/j.1440-1746.2000.02077.x.
35. Xiaolong C. A study on the association between drinking patterns and gastric Cancer and precancerous lesions. in Wuwei Cohort [D]; Lanzhou University; 2023.
36. Cao Y, Willett W C, Rimm E B, et al. Light to moderate intake of alcohol, drinking patterns, and risk of cancer: results from two prospective US cohort studies [J]. BMJ. 2015;351:h4238. https://doi.org/10.1136/bmj.h4238.
37. Rota M, Possenti I. Dose-response association between cigarette smoking and gastric cancer risk: a systematic review and meta-analysis [J]. Gastric Cancer. 2024;27(2):197–209. https://doi.org/10.1007/s10120-023-01459-1.
38. Hatta W, Koike T, Asano N, et al. The impact of tobacco smoking and alcohol consumption on the development of gastric cancers [J]. Int J Mol Sci. 2024;25(14). https://doi.org/10.3390/ijms25147854.
39. Yaofu F, Yanmin W. Analysis of risk factors associated with precancerous lesion of gastric cancer in patients from Eastern China: a comparative study [J]. Chin J Gastroenterol Hepatol. 2014;23(02):143–6.
40. Yang L, Yuanyuan N. Diagnosis and treatment of precancerous lesions of gastric cancer [J]. J Dig Oncology(Electronic Version). 2022;14(02):113–8.
41. Isakov V. Autoimmune gastritis studies and gastric cancer: true renaissance or bibliometric illusion [J]. World J Gastroenterol. 2024;30(32):3783–90. https://doi.org/10.3748/wjg.v30.i32.3783.
42. Bin Z, Xueqi F, Xiaolong Z, et al. A case-control study on risk factors of gastric cancer in four provinces of Huaihe river basin [J]. Chin J Prev Control Chronic Dis. 2022;30(06):437–41. https://doi.org/10.16386/j.cjpccd.issn.1004-6194.2022.06.008.
43. Linglin Z. Construction and verification of risk prediction model for gastric Cancer [D]. Chengdu University of TCM; 2023.
44. Eom B W, Joo J, Kim S, et al. Prediction model for gastric Cancer incidence in Korean population [J]. PLoS ONE. 2015;10(7):e0132613. https://doi.org/10.1371/journal.pone.0132613.
45. Zengyun W, Jing N, Yue W, et al. Research progress on risk factors of gastric cancer in families [J]. China Med. 2023;18(08):1264–7.
46. Shin C M, Yang K=. Stomach cancer risk in gastric cancer relatives: interaction between Helicobacter pylori infection and family history of gastric cancer for the risk of stomach cancer [J]. J Clin Gastroenterol. 2010;44(2):e34–9. https://doi.org/10.1097/MCG.0b013e3181a159c4.
47. Chen B, Wang Y, Tang W, et al. Association between PPARgamma, PPARGC1A, and PPARGC1B genetic variants and susceptibility of gastric cancer in an Eastern Chinese population [J]. BMC Med Genomics. 2022;15(1):274. https://doi.org/10.1186/s12920-022-01428-0.

Publisher's note