Dean&Francis

ISSN 2959-6157

Causal Discovery in Diabetes and its Complications

Xiaohan Pan[1, *]

[1] Sino-French Institute, Renmin University of China, Beijing, 100000, China

*Corresponding author: panxiaohan2023202533@ruc.edu.cn

Abstract:

Although previous studies have demonstrated that diabetes and its complications and other factors have high correlation, there is a lack of research using data-driven techniques to infer causality in this field. In this research, causal discovery is used to deal with a purified dataset containing 70,692 survey responses from the CDC's BRFSS 2015, provided on Kaggle. The final causal graph result obtained through causal discovery strongly aligns with the theory of metabolic syndrome. The findings indicate that abdominal obesity is a leading factor in the development of diabetes and hypertension, and this causal relationship is strong. *Hyperglycemia* significantly leads to both hypertension and hyperlipidemia, while hypertension also markedly exacerbates dyslipidemia. There is a strong causal relationship between each pair of these conditions. hyperglycemia, hypertension, and hyperlipidemia contribute to the development of cardiovascular disease, with a moderate strength of causal relationship; hyperglycemia, hypertension, and hyperlipidemia all contribute to the development of stroke, with a weak strength of causal relationship. Cardiovascular disease leads to stroke with a weak strength of causal relationship. Aging is strongly causally linked to hyperlipidemia, while its causal relationship with cardiovascular disease is moderate. Using these understandings of the causal relationships between diabetes and its complications and other factors, the author can enhance the health status and quality of life of patients with diabetes on multiple levels by implementing precise prevention strategies, optimizing treatment options, reducing healthcare costs, and improving public health policies.

Keywords: Diabetes; metabolic syndrome theory; causal discovery.

1. Introduction

Diabetes is a widespread metabolic disease with a rapidly increasing global prevalence, making it a major chronic condition that threatens public health. The International Diabetes Federation (IDF) reported that around 537 million adults (aged 20-79) were living with diabetes globally in 2021. This figure is expected to increase to 643 million by 2030 and 783 million by 2045 [1]. As the disease progresses, patients may develop various complications, including cardiovascular disease, stroke. These complications not only affect the quality of life of patients but also significantly increase the burden of the disease and healthcare cost [2, 3]. The rising incidence of diabetes and its associated complications underscores the urgent need for advanced methods to understand the complex relationships between them, therefore enhancing the ability to identify high-risk populations and to provide developing targeted prevention and treatment strategies [4]. This paper aims to uncover causal relationships between diabetes and its related factors.

Currently, the most prevailing theory explaining the pathogenesis of diabetes is metabolic syndrome theory. The pathophysiology of metabolic syndrome involves multiple complex mechanisms, many of which remain unclear. In addition to genetic and epigenetic factor, certain lifestyle factors and environmental factors (such as overeating and lack of physical exercise) are also considered major contributing factors to the development of metabolic syndrome. visceral obesity has been identified as a critical trigger that initiates several pathological pathways of metabolic syndrome [5]. Recent studies indicate that insulin resistance, chronic inflammation, and neurohormonal activation play a crucial role in the onset of metabolic syndrome and its advancement to cardiovascular diseases (CVD) and type 2 diabetes mellitus (T2DM) [5]. Current research on metabolic syndrome primarily focuses on clinical studies and traditional correlation analyses, with a lack of research using data-driven techniques for causal inference. By using tools and methods such as causal diagrams, structural equation models, and counterfactual inference, researchers can go beyond merely relying on randomized controlled trials (RCTs) to automatically identify and validate potential causal relationships from large datasets [6].

The study first conducts preliminary and further screenings of the variables in the dataset, using univariate logistic regression analysis and multivariate logistic regression analysis and ultimately selects 10 variables to construct causal graph [7]. To explore causal graph, model selection using BIC score and global search using the hill-climbing algorithm are employed, and random restarts is used to avoid the issue of being trapped in local optimum potentially caused by the hill-climbing algorithm [8-10]. To obtain a concise and interpretable causal graph, six disease variables are initially used to discover causal graph, yielding results consistent with the metabolic syndrome theory. Subsequently, the causal relationships among these six variables and the fact that gender and age have no parent nodes in the causal graph are introduced as background knowledge into causal discovery process to construct an initial causal graph containing 10 variables. Based on this, this paper uses neural networks to derive the structural equation model (SEM) through regression analysis [11]. Then, the paper uses the counterfactual inference (based on SEM) to show the strength of causal relationships between variables, remove neural network nodes with poor predictive performance and retain only causal relationships with non-zero counterfactual inference results, thereby arriving at the final causal graph. The final causal graph is found to be highly consistent with metabolic syndrome theory.

2. Methods

2.1 Data Source

In this paper, a dataset of 70,692 survey responses from the CDC's BRFSS 2015 that has been cleaned and shared by Alex Teboul on Kaggle, is used. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual health-focused telephone survey conducted by the CDC.

The data contains 22 variables as Table1 mentions. To process the data, this research converts all binary variables (such as high blood pressure, smoking, etc.) into factor type; category variables (such as age, BMI, income, etc.) into numeric types and then standardize them (table 1).

ISSN 2959-6157

Table 1. Name and explanation of variables

Variable	Logogram	Meaning
HighBP	x_1	0 = No hypertension, 1 = Hypertension
HighChol	x_2	0 = No hypercholesterolemia, 1 = Hypercholesterolemia
CholCheck	x_3	0 = Has not had a cholesterol check in the last 5 years, 1 = Has had a cholesterol check in the last 5 years
BMI	x_4	Body Mass Index
Smoker	x_5	0 = No, 1 = Yes (at least 100 cigarettes in entire life)
Stroke	x_6	History of stroke diagnosis: 0 = No, 1 = Yes
HeartDiseaseorAttack	x_7	Coronary heart disease (CHD) or myocardial infarction (MI) diagnosis: 0 = No, 1 = Yes
PhysActivity	x_8	Engaged in physical activity in the past 30 days (excluding work-related activities): 0 = No, 1 = Yes
Fruits	x_9	Fruit consumption (1 or more times per day): 0 = No 1 = Yes
Veggies	x_{10}	Vegetable consumption (1 or more times per day): 0 = No 1 = Yes
HvyAlcoholConsump	x_{11}	Heavy alcohol consumption (men: 14+ drinks per week, women: 7+ drinks per week): 0 = No, 1 = Yes
AnyHealthcare	x_{12}	Health insurance coverage: 0 = No, 1 = Yes
NoDocbcCost	x_{13}	Unable to see a doctor in the past 12 months due to cost: 0 = No, 1 = Yes
GenHlth	x_{14}	General health rating (1-5): 1 = Excellent, 2 = Very good, 3 = Good, 4 = Fair, 5 = Poor
MentHlth	x_{15}	Days of poor mental health in the past 30 days: 1-30
PhysHlth	x_{16}	Days of poor physical health in the past 30 days: 1-30
DiffWalk	x_{17}	Serious difficulty walking or climbing stairs: 0 = No, 1 = Yes
Sex	x_{18}	0 = female, 1 = male
Age	x_{19}	Age category 13-level :1 = 18-24, 9 = 60-64, 13 = 80 or older
Education	x_{20}	Education level scale 1-6: 1 = No schooling or kindergarten, 2 = Grades 1-8, 3 = Grades 9-11, 4 = High school graduate or GED, 5 = 1-3 years of college/technical school, 6 = College graduate (4+).
Income	x_{21}	Income scale 1-8: 1 = less than $10,000, 5 = less than $35,000, 8 = $75,000 or more
Diabetes_binary	Y	0 = no diabetes, 1 = prediabetes, diabetes

2.2 Protocol Statistical Analysis

The logistic regression model analysis is performed using R software. The bar chart, causal graph, neural network model, and counterfactual inference are all implemented and analyzed using Python. The causal graph is performed using the "pgmpy" package, and the neural network model is performed using the "sklearn" package. The algorithm flowchart and the construction of the causal graph are carried out using yEd software.

2.2.1 The logistic regression model

Logistic regression is a probabilistic nonlinear regression model used to predict the outcome of a binary dependent variable. It is widely used in the medical field to predict disease incidence. In univariate regression analysis, each independent variable is individually analyzed with the dependent variable by constructing a logistic regression model, which provides an initial assessment of which factors affect the dependent variable. Multivariate analysis, on the other hand, builds a logistic regression model with the selected variables and the dependent variable simultaneously, based on the univariate analysis. Multivariate

analysis can adjust for the influence of confounding factors, making the results more reliable. The formula for the Logistic Regression model is as follows:

$$ln\left(P/(1-P)\right)=\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_m X_m \quad (1)$$

Where β? is the intercept term, indicating the value of $ln(P/(1-P))$ when all the independent variables in the model are equal to 0. $ln(P/(1-P))$ is referred to as $logitP$, representing the natural logarithm of the probability P (this transformation avoids the issue where independent variables cannot form a linear relationship with a binary dependent variable). Since the value of probability P ranges between 0 and 1, logitP can take on values within an infinite range. β_m is the regression coefficient, indicating the change in $logitP$ for a one-unit variation in the independent variable X_m while keeping other variables fixed. In other words, they represent the relationship between the independent variable and the natural logarithm of the odds ratio (OR).

2.2.2 Causal discovery

Judea Pearl in the 1980s proposed methods for causal discovery based on Bayesian networks and causal diagrams. Causal discovery refers to the process of uncovering causal relationships between variables from data, rather than merely identifying traditional correlation relationships. Causal discovery can control for confounding variables and exclude reverse causality, helping us gain a deeper understanding of a system's causal mechanisms.

Causal diagrams are at the core of causal discovery. Graph structures, typically Directed Acyclic Graphs (DAGs), are used to represent causal relationships between variables. Each node represents a variable, and the edges represent causal relationships from the cause variable to the effect variable. The task of causal discovery is to identify the causal structure among variables from the data.

Algorithms for discovering causal diagrams are generally divided into two categories: algorithms based on conditional independence tests and algorithms based on scoring methods. The BIC (Bayesian Information Criterion) score-based algorithm is a commonly used method. BIC is a criterion for model selection that balances the model's goodness of fit with its complexity, penalizing overly complex models. The formula for BIC is as follows:

$$BIC=-2\times ln(L)+k\times ln(n) \quad (2)$$

L represents the likelihood of the model, k denotes the number of parameters, and n refers to the sample size.

2.2.3 Causal graph optimization

To further explore causal graph, the paper uses BIC score for model selection and the hill-climbing algorithm for global search and applies random restarts to effectively avoid the potential issues of being trapped in a local optimum, which may arise from the hill-climbing algorithm. The flowchart of the algorithm is shown in Figure 1.

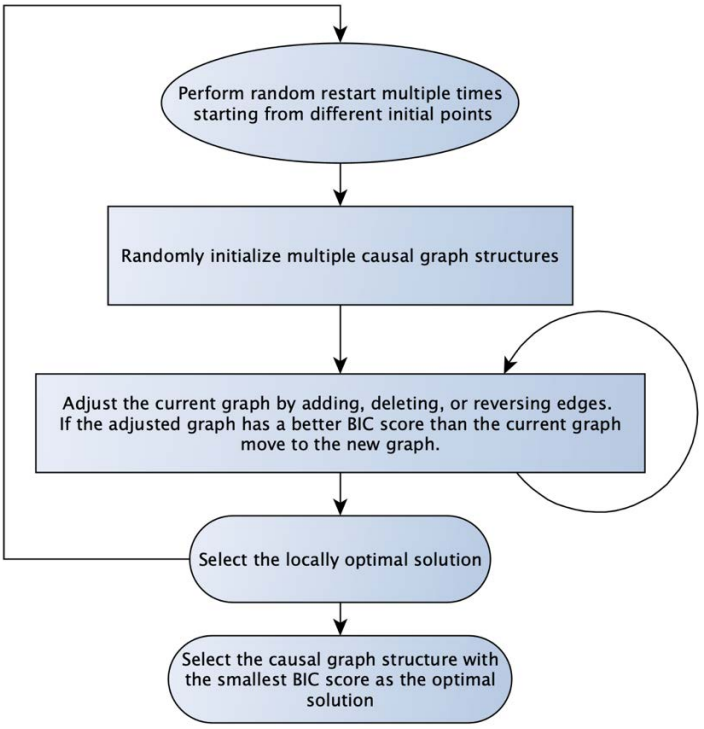

Fig. 1 The flowchart of algorithm for causal graph optimization

ISSN 2959-6157

2.2.4 Neural Network

Structural Equation Modeling (SEM) describes the complex quantitative relationships between variables in a causal diagram. To capture complex nonlinear relationships, this paper uses an extended SEM acquisition method. Specifically, a neural network is employed to perform regression analysis, fitting the nonlinear mapping between independent and dependent variables, thereby forming nonlinear structural equations. The algorithm flowchart is shown in Figure 2.

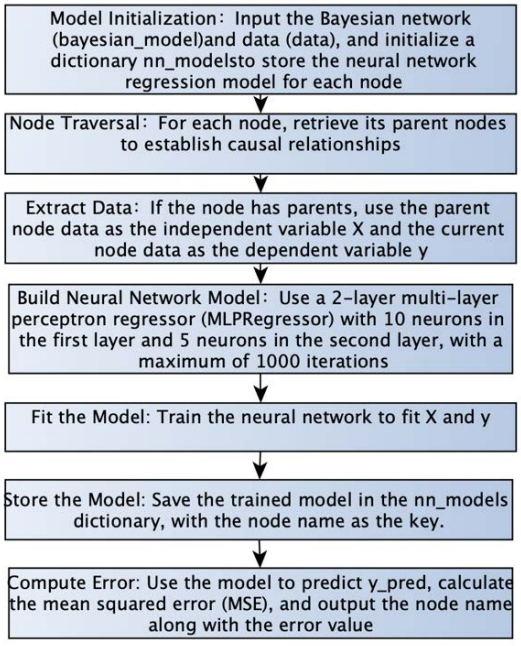

Fig. 2 The flowchart of algorithm for SEM

2.2.5 Counterfactual inference

Counterfactual Inference refers to the process of speculating "what would happen if the state of a certain variable were different from the actual situation." It helps answer "what if" questions, predicting or estimating changes in outcomes by altering certain known conditions. Based on the obtained structural equations, this paper assumes that the value of an independent variable differs from reality (e.g., changing a treatment variable from 0 to 1). This paper then modifies the corresponding structural equations and recalculate the values of other related variables to derive the results under the counterfactual scenario. Counterfactual inference can be used to demonstrate the strength of causal relationships between variables.

3. Result and Discussion

3.1 Select Variables

This study uses univariate logistic regression analysis to preliminarily screen the variables, analyzing each variable in conjunction with Diabetes_binary. The results show that the p-values of all 21 variables in the data were less than 0.05, indicating the need for further screening. The study uses multivariate logistic regression for re-screening the variables, analyzing the 21 variables together in conjunction with Diabetes_binary. To simplify the results, this study selects variables with a p-value < 0.001, indicating high significance. The results in Figure 3 show that HighBP, HighChol, CholCheck, Stroke, HeartDiseaseorAttack, HvyAlcoholConsump, GenHlth, Sex, MentHlth, PhysHlth, DiffWalk, Age, Income, BMI are highly significant.

CholCheck represents frequency of blood lipid tests, which, according to common knowledge, is unrelated to the pathogenesis of diabetes. DiffWalk, GenHlth, MentHlth, PhysHlth are the respondents' subjective judgment of their health status, which cannot accurately reflect objective reality. After removing these five variables, the study ultimately selects 10 variables including Diabetes_binary, HighBP, HighChol, Stroke, HeartDiseaseorAttack, HvyAlcoholConsump, Sex, Age, Income, BMI to construct causal graph. Figure 3 presents the descriptive statistics of the final selected variables. Apart from the balanced Diabetes_binary, the variables HighBP, HighChol, Sex are relatively balanced, while the other variables show significant skewness. Notably, Stroke, HeartDiseaseorAttack, HvyAlcoholConsum are highly skewed, which could potentially impact the final results.

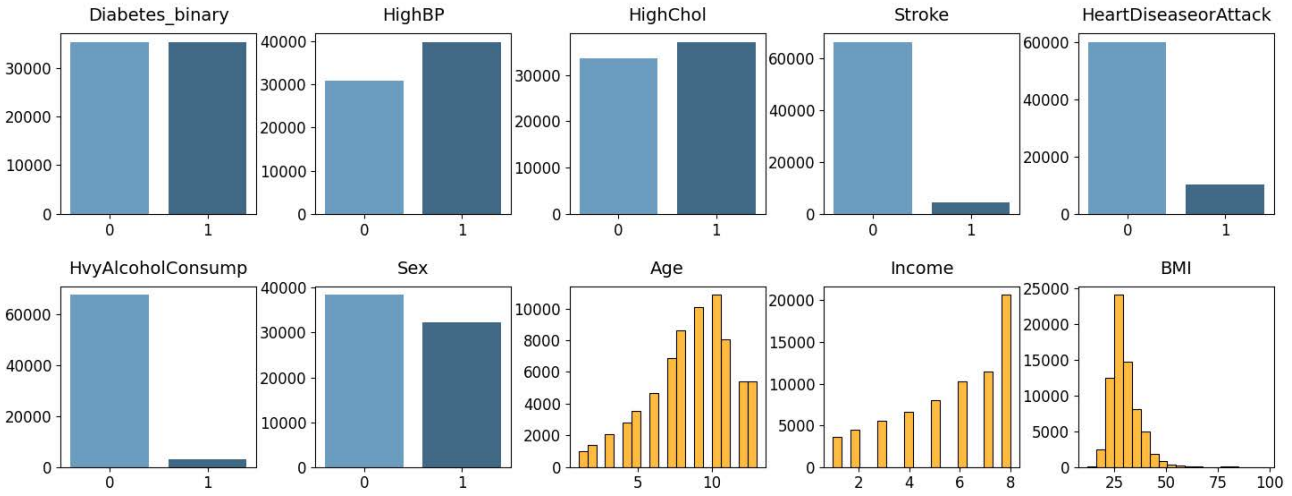

Fig. 3 Bar chart of the final selected variables

3.2 Find Preliminary Causal Graph

To discover causal graph, this study first applies an algorithm to 10 variables. This initially results in the outcome whose direction of causal edges is confused and whose BIC score is too high. This suggests that this causal graph is overly complex and the data fit is poor. To achieve a simpler and more effective result, considering the intricate relationships between diabetes and its complications, this study applies the algorithm to discover causal graph among 6 disease-related variables out of the 10 first. These disease variables are Diabetes_binary, HighBP, HighChol, Stroke, HeartDiseaseorAttack, BMI (an indicator variable for abdominal obesity). The results are shown in Figure 4, as the relationships among the blue variables. Abdominal obesity leads to hyperglycemia and hypertension, hyperglycemia leads to hypertension and hyperlipidemia, and hypertension leads to hyperlipidemia; both hypertension and hyperlipidemia lead to heart disease and stroke, and heart disease leads to stroke. This aligns with the prevailing theory of the pathogenesis of diabetes and its accompanying complications-metabolic syndrome theory.

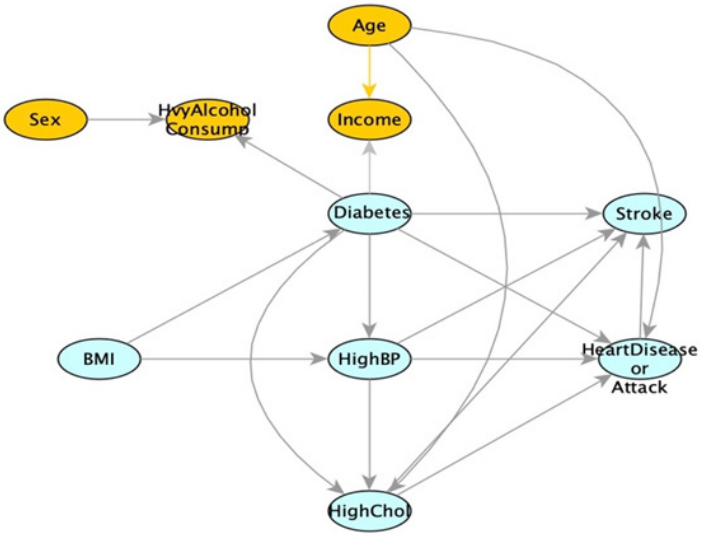

Fig. 4 The preliminary causal graph

ISSN 2959-6157

Based on the plausible causal relationships among the disease variables, this study then adds 4 more variables and applies the algorithm to discover causal graph among the 10 variables. Specifically, the paper incorporates the causal relationships among six variables, as well as the background knowledge that gender and age have no parent nodes in the causal graph, into the causal discovery process. The preliminary causal graph is concise and consistent with common sense.

3.3 Obtain Causal Relationship Strength

Having obtained the qualitative causal relationship, this study also aims to obtain the quantitative causal relationship. Therefore, this study needs to derive the structural equation model (SEM) based on the causal graph results. To capture complex nonlinear relationships, this paper uses neural networks to obtain the structural equation model (SEM) through regression analysis. The neural network regression model shows that, except for the MSE of node Income being 0.9286, the MSE of the remaining nodes is between 0.0405 and 0.2232. A high MSE indicates poor predictive performance at node Income, so this study chooses to remove node Income.

Then, the paper uses the counterfactual inference (based on SEM) to show the strength of causal relationships between variables, removes neural network nodes with poor predictive performance and retains only causal relationships with non-zero counterfactual inference results, thereby arriving at the final causal graph. The final causal graph is found to be highly consistent with metabolic syndrome theory. The final causal graph is shown in Figure 5.

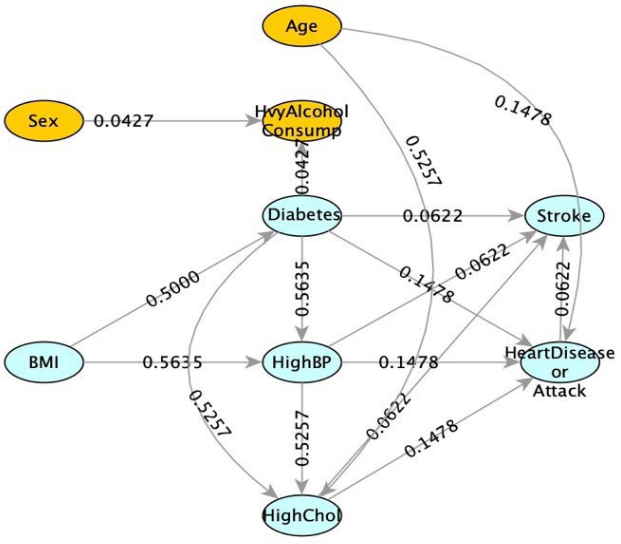

Fig. 5 The final causal graph

4. Conclusion

The final causal graph result obtained through causal discovery is strongly aligns with the theory of metabolic syndrome. The results indicate that abdominal obesity is a significant contributor to diabetes and hypertension. Hyperglycemia significantly leads to both hypertension and hyperlipidemia, while hypertension also markedly exacerbates dyslipidemia. There are strong causal relationships between each pair of these conditions.

According to the pathological mechanism of metabolic syndrome (MetS), abdominal obesity is a crucial trigger for initiating various pathological pathways of MetS, as it is often accompanied by insulin resistance-a common underlying factor in many diseases associated with MetS. Insulin resistance impairs the inhibition of lipolysis in adipose tissue, leading to elevated levels of free fatty acids (FFA) in circulation. These FFAs promote gluconeogenesis and lipogenesis in the liver while inhibiting glucose uptake in muscles, ultimately resulting in hyperglycemia. FFAs can also induce oxidative stress, increasing reactive oxygen species (ROS) and reducing nitric oxide (NO) levels, leading to vasoconstriction and the onset of hypertension. Additionally, FFAs promote the synthesis of very-low-density lipoprotein (VLDL) in the liver and activate cholesterol ester transfer protein (CETP), facilitating the transfer of triglycerides from very low-density lipoprotein (VLDL) to high-density lipoprotein (HDL), which

in turn reduces HDL levels and contributes to hyperlipidemia. The increase in FFAs, in turn, alters the insulin signaling cascade in various organs, worsening insulin resistance and creating a vicious cycle that exacerbates hyperglycemia, hypertension, and hyperlipidemia.

The results of this study show that insulin resistance induced by abdominal obesity first leads to hyperglycemia and hypertension. The FFA increase associated with hyperglycemia, coupled with the vicious cycle of aggravated insulin resistance, further exacerbates hypertension. The increase in FFAs related to both hypertension and hyperglycemia, along with the worsening of insulin resistance, ultimately leads to hyperlipidemia. Each layer of this causal relationship demonstrates a strong association.

Hyperglycemia, hypertension, and hyperlipidemia contribute to the development of cardiovascular disease, with a moderate strength of causal relationship; hyperglycemia, hypertension, and hyperlipidemia all contribute to the development of stroke, with a weak strength of causal relationship. Cardiovascular disease leads to stroke with a weak strength of causal relationship.

Hyperglycemia and insulin resistance can also damage the vascular endothelium by increasing oxidative stress and inflammation. Hypertension contributes to endothelial damage and arteriosclerosis by increasing the pressure on blood vessel walls. Hyperlipidemia further promotes the formation of atherosclerosis. When these damages occur in the cerebral blood vessels, they can lead to stroke; when they occur in the cardiovascular system, they can result in heart disease. From a pathological perspective, the heart's blood vessels, particularly the coronary arteries, are more susceptible to disease than the cerebral vessels. The coronary arteries branch directly from the aorta and are subject to higher pressure, making them more vulnerable to the effects of hypertension. Additionally, their smaller diameter makes them more prone to narrowing or obstruction caused by atherosclerotic plaques. In contrast, cerebral blood vessels have multiple circulatory pathways (such as the anterior cerebral artery and the Circle of Willis) that can compensate for the pathology of a single vessel to some extent. Cardiovascular thrombi entering the brain through the bloodstream can cause stroke, but this occurrence is relatively rare.

Aging is strongly causally linked to hyperlipidemia, while its causal relationship with cardiovascular disease is moderate. This is because, with increasing age, the metabolic rate in the body generally declines, which can affect lipid metabolism. Additionally, aging naturally leads to the deterioration of blood vessels.

The study also shows that being male and having hyperglycemia may contribute to alcohol consumption habits, though the causal relationship is relatively weak. This weak explanatory power could be due to a high bias in the variable related to alcohol consumption.

References

[1] International Diabetes Federation. IDF Diabetes Atlas, 10th edn. Brussels, Belgium: International Diabetes Federation, 2021.

[2] Andersson E, et al. Costs of diabetes complications: hospital-based care and absence from work for 392,200 people with type 2 diabetes and matched control participants in Sweden. Diabetologia, 2020, 63: 2582-2594.

[3] Gruss SM, Nhim K, Gregg E, et al. Public Health Approaches to Type 2 Diabetes Prevention: the US National Diabetes Prevention Program and Beyond. Curr Diab Rep, 2019, 19: 78.

[4] Fahed G, Aoun L, Bou Zerdan M, et al. Metabolic syndrome: updates on pathophysiology and management in 2021. International Journal of Molecular Sciences, 2022, 23(2).

[5] Pearl J. Causality: Models, Reasoning, and Inference. 2nd ed. New York: Cambridge University Press, 2009.

[6] Bursac Z, Gauss CH, Williams DK, et al. Purposeful selection of variables in logistic regression. Source Code Biol Med, 2008, 3: 17

[7] Gámez J A, Mateo J L, Puerta J M. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. Data Min Knowl Disc, 2011, 22: 106-148.

[8] Tsamardinos I, Brown L E, Aliferis C F. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn, 2006, 65: 31-78.

[9] Charnay C, Lachiche N, Braud A. Construction of Complex Aggregates with Random Restart Hill-Climbing. In: Davis J, Ramon J, eds. Inductive Logic Programming. Lecture Notes in Computer Science, vol 9046. Springer, Cham, 2015.

[10] Zaidan A S, Arianpoor A. Artificial Neural Networks and Structural Equation Modeling: Marketing and Consumer Research Applications. Springer, 2023.

[11] Rastogi S. Structural equation model (SEM)-neural network (NN) model for predicting quality determinants of e-learning management systems. Electronics, 2017.