Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 5552085, 12 pages
https://doi.org/10.1155/2021/5552085

Research Article

Improved Functional Causal Likelihood-Based Causal Discovery Method for Diabetes Risk Factors

Xiue Gao,[1] Wenxue Xie,[2] Zumin Wang,[2] Bo Chen ,[1] and Shengbin Zhou[1]

[1]*College of Information Engineering, Lingnan Normal University, Guangdong 524048, China*
[2]*College of Information Engineering, Dalian University, Dalian 116622, China*

Correspondence should be addressed to Bo Chen; chenbo20040607@126.com

Received 19 January 2021; Revised 26 April 2021; Accepted 4 May 2021; Published 17 May 2021

Academic Editor: John Mitchell

Diabetes mellitus is a disease that has reached epidemic proportions globally in recent years. Consequently, the prevention and treatment of diabetes have become key social challenges. Most of the research on diabetes risk factors has focused on correlation analysis with little investigation into the causality of these risk factors. However, understanding the causality is also essential to preventing the disease. In this study, a causal discovery method for diabetes risk factors was developed based on an improved functional causal likelihood (IFCL) model. Firstly, the issue of excessive redundant and false edges in functional causal likelihood structures was resolved through the construction of an IFCL model using an adjustment threshold value. On this basis, an IFCL-based causal discovery algorithm was designed, and a simulation experiment was performed with the developed algorithm. The experimental results revealed that the causal structure generated using a dataset with a sample size of 2000 provided more information than that produced using a dataset with a sample size of 768. In addition, the causal structures obtained with the developed algorithm had fewer redundant and false edges. The following six causal relationships were identified: insulin→plasma glucose concentration, plasma glucose concentration→body mass index (BMI), triceps skin fold thickness→BMI and age, diastolic blood pressure→BMI, and number of times pregnant→age. Furthermore, the reasonableness of these causal relationships was investigated. The algorithm developed in this study enables the discovery of causal relationships among various diabetes risk factors and can serve as a reference for future causality studies on diabetes risk factors.

1. Introduction

With the steady increase in the number of diabetic patients worldwide, diabetes mellitus has become the third most serious threat to human health after cerebro-cardiovascular diseases and malignant tumours [1]. Diabetes is a chronic metabolic disorder that can be caused by a wide variety of risk factors. It leads to disturbances in fat and protein metabolism, resulting in chronic injury or failure of multiple organs [2]. Diabetes severely impacts human health and imposes a heavy burden on families and societies; hence, there is a pressing need for effective prevention and treatment of diabetes. The analysis of the relationships among various risk factors and between diabetes and risk factors is essential to elucidate the pathogenesis of diabetes and is a precondition for diabetes prevention and treatment. Previous research in China and other countries has largely focused on two areas: (1) the analysis of risk factors for diabetes onset and (2) the construction of prediction models for diabetes onset.

(1) Research on the analysis of risk factors for diabetes onset primarily comprises two activities: the exploration of new risk factors and relationship analysis of risk factors. The investigation of new risk factors enables the discovery of potential factors for diabetes onset, which is beneficial for understanding diabetes aetiologies and may facilitate the effective prevention of diabetes. As the pathogenesis of diabetes involves multiple factors, the analysis of the relationships among these risk factors is particularly important and of practical and clinical significance. (i) Researchers have discovered many new risk factors of diagnostic and predictive significance. For instance, Fizelova et al. [3] found that the

apolipoprotein B/LDL cholesterol ratio and apolipoprotein A1/HDL cholesterol ratio are the strongest predictors of the worsening of glycaemia and incidence of type 2 diabetes, respectively, in Finnish men. Lankinen et al. [4] identified plasma fatty acids as a potential predictor for glycaemia and a risk factor for type 2 diabetes mellitus (T2DM) in Finnish men. Further, Yazdanpanah et al. [5] found that glycated albumin (GA) provides more accurate diabetes diagnosis than glycated haemoglobin. Another study by Huang et al. [6] revealed that adiponectin (ADPN) combined with fibroblast growth factor 21 (FGF-21) and adipocyte fatty acid binding protein (A-FABP) are of great clinical significance in the early diagnosis and risk prediction of T2DM and could serve as key markers for the prediction of T2DM onset in high-risk populations. Bellia et al. [7] demonstrated the clinical usefulness of GA in the diagnosis of diabetes in a high-risk Caucasian population. In another study, Tatsukawa et al. [8] found that the risk of diabetes in the Japanese population was significantly positively correlated with trunk fat and significantly negatively correlated with leg fat. Li et al. [9] revealed that the age of alcohol onset and drinking duration are risk factors for T2DM. (ii) Studies on relationship among the various risk factors have provided a basis and direction for the investigation of potential aetiologies of diabetes. Zhao et al. [10] explored the correlations of trace elements in serum with serum glucose and body composition indicators in T2DM patients and concluded that the correction of trace element metabolism disorders in T2DM patients may be of great significance for diabetes treatment and the prevention of complications. Tillin et al. [11] revealed that branched chain and aromatic amino acids, particularly tyrosine, may be potential treatment targets for diabetes in South Asian populations. In addition, Cui and Feng [12] found that body mass index (BMI) is positively correlated with body fat percentage and abdominal-glute ratio, which indicates that body fat percentage may be clinically significant for diabetes diagnosis. Huang et al. [13] constructed a correlation network with biomarkers related to T2DM, which showed that the leptin system plays a key role in diabetes development. Meanwhile, Zhu et al. [14] studied the relationship between diabetes and body composition and found that visceral fat content, total fat content, total lean body mass, trunk lean mass, and limb lean mass are influencing factors of glycated haemoglobin. Therefore, glycaemic control in T2DM patients may be associated with lean body weight. Through Mendelian randomisation analysis, Liu et al. [15] found that there is a causal relationship between the genetically driven nonalcoholic fatty liver disease (NAFLD) and central obesity, both of which are risk factors for diabetes

(2) Early research on diabetes prediction models mainly involved the use of statistical regression methods for model construction, with typical examples including a model developed by Chien et al. for predicting T2DM risk in the Taiwanese population [16], a prediction model for diabetes onset developed by Li et al. [17], a classification tree model for diabetes prediction in rural Chinese [18], a model for the prediction of T2DM risk in Japanese Americans [19], the Finnish Diabetes Risk Score tool [20], and a diabetes risk prediction model for a mixed African American and non-Hispanic white population [21]. In recent years, rapid developments in artificial intelligence techniques have led to the adoption of machine learning methods to construct diagnostic and predictive models of various diseases. Intelligent diagnosis and prediction methods for different diseases can be classified into two categories: one based on traditional single learner and the other based on multiple learners, such as the diabetes diagnosis method based on a single learner proposed by Rahman et al. [22] and the congestive heart failure diagnosis method based on multiple learners proposed by Isler et al. [23]. In the diagnosis and prediction of diabetes mellitus, the approach based on a single learner can provide satisfactory results with higher efficiency. For instance, Wang and Chen [24] utilised a support vector machine (SVM) with different kernel functions to construct prediction models for T2DM risk and found that the radial basis function-based SVM model provided the best predictive effects. Song et al. [25] and Chen et al. [26] reported the application of back-propagation neural network models to T2DM risk prediction. In addition, some researchers have improved the traditional single learner approach for better diagnosis and prediction. Erkaymaz et al. [27] found that Newman-Watts small-world feedforward neural networks have better accuracy in diagnosing diabetes, by comparing two different small-world feedforward neural networks. Geman et al. [28] used an adaptive neuro-fuzzy inference method to establish a diabetes classification and prediction system, which provided good classification and prediction accuracy. Further, several scholars have committed to exploring diabetes prediction methods based on multiple learners for better accuracy. For example, Liu et al. [29] developed a diabetes prediction model through the integration of SVM and the random forest (RF) technique and found that the integrated model provided superior classification performance compared with single classifiers. López et al. [30] used the RF technique to identify single-nucleotide polymorphisms in T2DM and to construct a decision-support tool for diabetes risk prediction. Wu et al. [31] used deep neural network and logistic regression models to predict gestational diabetes in the Chinese population, with better prediction performance that previous methods

Research on relationship among risk factors may enable the discovery of previously unknown physiological and

pathological phenomena of diabetes, providing a theoretical basis for the elucidation of diabetes pathogenesis. However, existing studies on the relationships among risk factors mostly reflect the correlations rather than causality among these factors. Although diabetes prediction models are beneficial for diabetes prevention and early diagnosis, they are fundamentally statistical correlation models that do not reflect causality. Therefore, there is a pressing need for studies on the causality of diabetes risk factors, as the determination of the pathological and physiological causal relationships of diabetes is of great theoretical significance and could provide clinical guidance for diabetes prevention and treatment.

Randomised controlled trials (RCTs) [32] constitute a traditional method of causality discovery. However, substantial interventions are required for the experimental group in an RCT, which are costly and may entail ethical and moral violations. These issues can be avoided by using observational data-based causal discovery methods, but noise in the data may influence the effects of causal discovery algorithms. In situations with significant noise, functional causal likelihood- (FCL-) based algorithms [33] can effectively discover causal relationships. However, in the discovery of causal relationships among diabetes risk factors, numerous redundant and erroneous causal edges are generated when using these algorithms. To overcome this problem, we developed an improved functional causal likelihood- (IFCL-) based diabetes risk factor causal discovery algorithm to uncover causal relationships among diabetes risk factors. Our study is the first to use the causal discovery algorithm to explore the causal relationship between diabetes risk factors.

The contributions of the present study are as follows:

(1) An IFCL model was developed by incorporating an adjustment threshold value α, which reduces the number of redundant and erroneous edges in the diabetes risk factor causal structures

(2) An IFCL-based diabetes risk factor causal discovery algorithm was subsequently constructed and used to generate optimised diabetes risk factor causal structures

(3) A simulation experiment was performed for comparative analysis of causal structures generated using different methods and sample sizes, and the significance of the identified causal relationships was assessed

The remainder of this paper is organised as follows. Section 2 provides the details of the IFCL model and diabetes risk factor causal discovery algorithm. Section 3 describes the experimental process and provides an analysis and discussion of the experimental results. Finally, Section 4 presents the study conclusions.

2. Materials and Methods

2.1. IFCL Model. The fundamental concepts of the FCL model are the assumption that the noise term is independent and is incorporated into the likelihood and that the likelihood over observational data is converted into the likelihood over the noise of the observational data and subsequently solved. Let $\{X_1, X_2, \cdots, X_N\}$ denote the variable set for diabetes risk factors, where N is the number of risk factor variables. G denotes the causal graph of the subset $X = \{X_1, X_2, \cdots, X_n\}$, $P(X_i = x)$ is the probability that $X_i = x$, and $P(X_i \mid P_i)$ indicates the probability of observations on X_i with conditions on the values of all its parents P_i, with $1 \le i \le n \le N$. Given that G satisfies the causal Markov condition [32, 34] and causal faithfulness condition [32], the joint distribution $P(X)$ can be expressed as follows:

$$P(X) = \prod_{i=1}^{n} P\left(X_i \mid X_{P_i}\right), \tag{1}$$

where X_{P_i} includes all parents of X_i. Given a group of observational data $\boldsymbol{O} = \{\overrightarrow{\boldsymbol{o_1}}, \overrightarrow{\boldsymbol{o_2}}, \cdots, \overrightarrow{\boldsymbol{o_j}}, \cdots, \overrightarrow{\boldsymbol{o_m}}\}$, where $\overrightarrow{\boldsymbol{o_j}}$ is an n-dimensional vector (i.e., $\overrightarrow{\boldsymbol{o_j}} = (o_{j,1}, o_{j,2}, \cdots, o_{j,n})$, $1 \le j \le m$), o_{j,P_i} can be used to denote the subvector of $\overrightarrow{\boldsymbol{o_j}}$ containing the observational values of X_{P_i}. By combining $P(X)$ and G, the log-likelihood of the observational data can be expressed as follows:

$$L(G; O) = \sum_{j=1}^{m} \sum_{i=1}^{n} \log\left(P\left(X_i = o_{j,i} \mid X_{P_i} = o_{j,P_i}\right)\right). \tag{2}$$

A search for causal networks by maximising the likelihood calculated using Equation (2) may not return true causality structures owing to the possible existence of different graphical structures providing exactly the same likelihood, which are known as Markov equivalence classes. To overcome the issues associated with Markov equivalence classes, it is necessary to introduce the concepts of causal function and noise.

Figure 1 shows a partial causal structure, with E_i and X_{P_i} denoting the randomised noise corresponding to X_i and the causal variable of X_i, respectively. An additive noise model $X_i = F_i(X_{P_i}) + E_i$ is adopted as the causal mechanism, with F_i being the causal function of X_i and the randomised noise variable E_i being independent of the causal variable X_{P_i}. Therefore, the following equation can be derived:

$$\begin{aligned} P\left(X_i = o_{j,i} \mid X_{P_i} = o_{j,P_i}\right) &= P\left(E_i = o_{j,i} - F_i\left(o_{j,P_i}\right) \mid X_{P_i}\right) \\ &= P\left(E_i = o_{j,i} - F_i\left(o_{j,P_i}\right)\right). \end{aligned} \tag{3}$$

From Equations (2) and (3), it can be seen that the likelihood over the observational data is equivalent to the likelihood over the noise of the observational data. Let $S = \langle G, F \rangle$ denote the causal structure. The likelihood over the noise of the observational data can then be obtained as follows:

$$L(S; O) = \sum_{j=1}^{m} \sum_{i=1}^{n} \log\left(P\left(E_i = o_{j,i} - F_i\left(o_{j,P_i}\right)\right)\right). \tag{4}$$

Equation (4) shows the converted target function. For

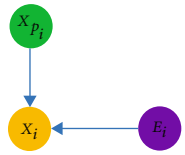

Figure 1: Partial causal structure consisting of X_{P_i} and X_i.

datasets with limited sample sizes, the equation must be regularised to avoid the generation of excessive redundant causal edges. By introducing the Bayesian information criterion penalty, the regularised likelihood can be expressed as follows:

$$L_B(S;O) = \sum_{i=1}^{n}\left(\sum_{j=1}^{m}\log\left(P\left(E_i = o_{j,i} - F\left(o_{j,P_i}\right)_i\right)\right) - \frac{d_i \log(m)}{2}\right). \tag{5}$$

Equation (5) represents the FCL model, with d_i being the number of coefficients used to estimate X_i. By maximising Equation (5), the causal graph structure can be obtained, i.e., $\max L_B(S;O) = \max_G \sup_F L_B(\langle G, F\rangle; O)$. This represents the solution process of the FCL-based causal discovery algorithm, which involves two steps: (1) generation of initial causal graphs by fitting and optimising the causal function $\sup_F L_B(\langle G, F\rangle; O)$; (2) searching for the causal graph with the maximum likelihood $\max_G L_B(\langle G, F\rangle; O)$ using the hill-climbing algorithm, with the local updating rule for X_i given by the following equation:

$$L'_{Bi}(S;O) = \sum_{j=1}^{m}\log\left(P\left(E_i = o_{j,i} - F_i\left(o_{j,P_i}\right)\right)\right) - \frac{d_i \log(m)}{2}. \tag{6}$$

The FCL of diabetes risk factors obtained after iteration is denoted as $L_B^*(S;O)$. As the termination condition for the hill-climbing algorithm in the search for the causal graph with the maximum target likelihood is $L_B^*(S;O) > L_B(S;O)$, where $L_B(S;O)$ is the FCL of the initial causal structure, excessive redundant or erroneous edges are present in the generated diabetes risk factor causal structures. Therefore, an adjustment threshold value is introduced into Equation (5) for correction, resulting in the following corrected model:

$$\bar{L}_B(S;O) = \sum_{i=1}^{n}\left(\sum_{j=1}^{m}\log\left(P\left(E_i = o_{j,i} - F_i\left(o_{j,P_i}\right)\right)\right) - \frac{d_i \log(m)}{2} + \alpha\right). \tag{7}$$

Equation (7) represents the modified diabetes risk factor IFCL model, with α being the adjustment threshold value. In the hill-climbing algorithm, Equation (6) remains the local updating rule for X_i, whereas the termination condition becomes $L_B^*(S;O) > \bar{L}_B(S;O)$. The likelihood without updated nodes during the iteration process is given by the following equation:

$$L_{Bi}(S;O) = \sum_{j=1}^{m}\log\left(P\left(E_i = o_{j,i} - F_i\left(o_{j,P_i}\right)\right)\right) - \frac{d_i \log(m)}{2} + \alpha. \tag{8}$$

The diabetes risk factor FCL of the kth iteration can be expressed as

$$L_B^*(S;O) = \sum_{i=1}^{n}\left(\sum_{j=1}^{m}\log\left(P\left(E_i = o_{j,i} - F_i\left(o_{j,P_i}\right)\right)\right) - \frac{d_i \log(m)}{2}\right) + \alpha_k \tag{9}$$

where α_k is the total threshold of the kth iteration. It can be seen from Equation (7) that the total threshold of the initial IFCL model is $n\alpha$, which can be regarded as the likelihood of each causal node increasing by the threshold α, namely, $L_{Bi}(S;O) = \sum_{j=1}^{m}\log\left(P\left(E_i = o_{j,i} - F_i\left(o_{j,P_i}\right)\right)\right) - \left(d_i \log(m)/2\right) + \alpha$. After each iteration, the likelihood of updating the node will decrease by α, and the total threshold will continue to decrease, namely, $\alpha_k < \alpha_l, k > l$. Therefore, a causal node with greater likelihood must be searched for in the iteration process to reach the iteration termination condition $L_B^*(S;O) > \bar{L}_B(S;O)$, which is the fundamental reason why the IFCL-based diabetes risk factor causal discovery algorithm can output a more optimised causal structure.

2.2. IFCL-Based Diabetes Risk Factor Causal Discovery Algorithm. Figure 2 shows a flowchart of the IFCL-based diabetes risk factor causal discovery algorithm. The detailed steps of the algorithm are as follows.

Step 1. The observational data for diabetes risk factors $\boldsymbol{O} = \{\overrightarrow{\boldsymbol{o_1}}, \overrightarrow{\boldsymbol{o_2}}, \cdots, \overrightarrow{\boldsymbol{o_j}}, \cdots, \overrightarrow{\boldsymbol{o_m}}\}$ are input into the algorithm and subjected to pretreatment and normalisation.

Step 2. Firstly, the regression method is adopted to estimate the causal function F_i corresponding to the causal edges. Next, the norm of the residual (noise) is calculated by regression. Kernel density estimation is subsequently employed to approximate the noise distribution to obtain the optimised causal function F_i, which is then used to generate the initial causal graph G.

Step 3. The likelihood over noise $\bar{L}_B$ is initialised using Equation (7), and $L_B{}^*$ is set to zero.

Step 4. The hill-climbing algorithm is used to search for the optimal causal graph. During each iteration, the addition, deletion, or reversion operation is performed on a single causal edge in G. The causal function F_i and causal graph are updated, and the updated causal graph is stored in G^*.

Step 5. G^* and G are compared, and the updating of local likelihoods is performed for nodes with changes using Equation (6) to obtain L'_{Bi}. The updated likelihoods $\sum_i L'_{Bi}$ and

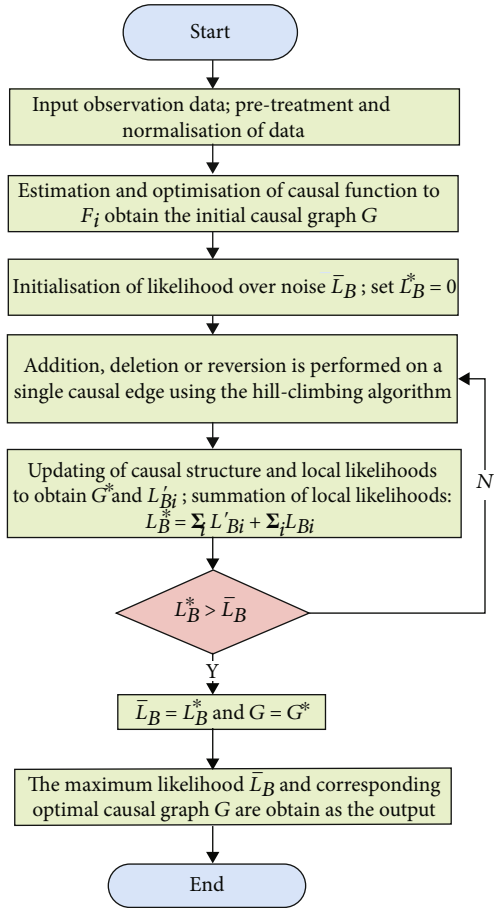

Figure 2: Flowchart of the IFCL-based diabetes risk factor causal discovery algorithm.

nonupdated likelihoods $\sum_i L_{Bi}$ are summed to obtain

$$L_B^* = \sum_i L'_{Bi} + \sum_i L_{Bi}. \tag{10}$$

Step 6. L_B^* and L_B are compared. If $L_B^* > \bar{L}_B$, then $\bar{L}_B = L_B^*$ and $G = G^*$, and the algorithm proceeds to Step 7. Otherwise, Step 4 is executed.

Step 7. The maximum likelihood $\bar{L}_B$ and corresponding optimal causal graph G are obtained as the output.

3. Results and Discussion

3.1. Experimental Data and Environment. Diabetes datasets with sample sizes of 768 (denoted as the $M = 768$ dataset) and 2000 (denoted as the $M = 2000$ dataset), which were obtained from the National Institute of Diabetes and Digestive and Kidney Diseases in the U.S.A. and Hospital Frankfurt in Germany, respectively, were downloaded from Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database; https://www.kaggle.com/chirag9073/diabetes-using-deep-learning/data) and used as the experimental data for this study. All subjects in the datasets were at least 21 years old. The datasets consisted of nine variables: number of times pregnant, plasma glucose concentration at 2 h in an oral glucose tolerance test, diastolic blood pressure (mmHg), triceps skin fold thickness (mm), 2 h serum insulin (muU/ml), BMI, diabetes pedigree function, age, and class variable for diabetes diagnosis. In particular, the diabetes pedigree function contains genetic information regarding diabetes history in the family of the subject. Except for the class variable, all other variables were subjected to causality analysis in this study. To maximise the retention of information, mean imputation was adopted to replace the missing values in the datasets. Z -score standardisation was performed on the raw data, and abnormal values were replaced by mean values.

The simulation experiment was carried out in the RStudio environment, and the program was written in R language. The computer used had an Intel (R) Core (TM) i7-6500U CPU with main frequency 2.50 GHz and 8 GB of RAM.

3.2. Experimental Results

3.2.1. Scatter Plots and Correlation Coefficients of Variable Pairs. To understand their correlation and provide a basis for subsequent experiments to analyse their causality, the scatter plots and correlation coefficients of variable pairs among the eight variables were generated for the $M = 768$ dataset (Figure 3) and $M = 2000$ dataset (Figure 4).

Figures 3 and 4 show scatter plots of the variable pairs in the bottom left corner, bar charts for each variable on the diagonal line from top left to bottom right, and correlation coefficients of the variable pairs in the top right corner. Figures 3 and 4 both show the scatter plots and correlation coefficients of 28 variable pairs. There are seven variable pairs with correlation coefficients less than 0.1 in Figure 3, while there are eight pairs of such cases in Figure 4.

In general, if the correlation coefficient of two variables is between 0 and 0.1, the relationship between the variables can be considered nonlinear. Therefore, variable pairs with correlation coefficients < 0.1 were discarded. Tables 1 and 2 show the variable pairs with correlation coefficients ≥ 0.1 and the corresponding P values. All P values are less than 0.01, which indicates the existence of significant linear relationships in the variable pairs.

3.2.2. Results of FCL-Based Causal Discovery. To better demonstrate and analyse the causal structure of diabetes risk factors, we set no. of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, insulin, BMI, age, and diabetes pedigree function to the variables X_1, X_2, X_3, X_4, X_5, X_6, X_7, and X_8, respectively.

To investigate the presence or absence of causality among the eight variables, a causal discovery experiment was performed with the $M = 768$ and $M = 2000$ datasets using an FCL-based causal discovery algorithm reported previously [33]. Figures 5 and 6 depict the resultant causal structures (named structures 1 and 2) and show 7 and 8 pairs of causal relationships, respectively. In Figures 5

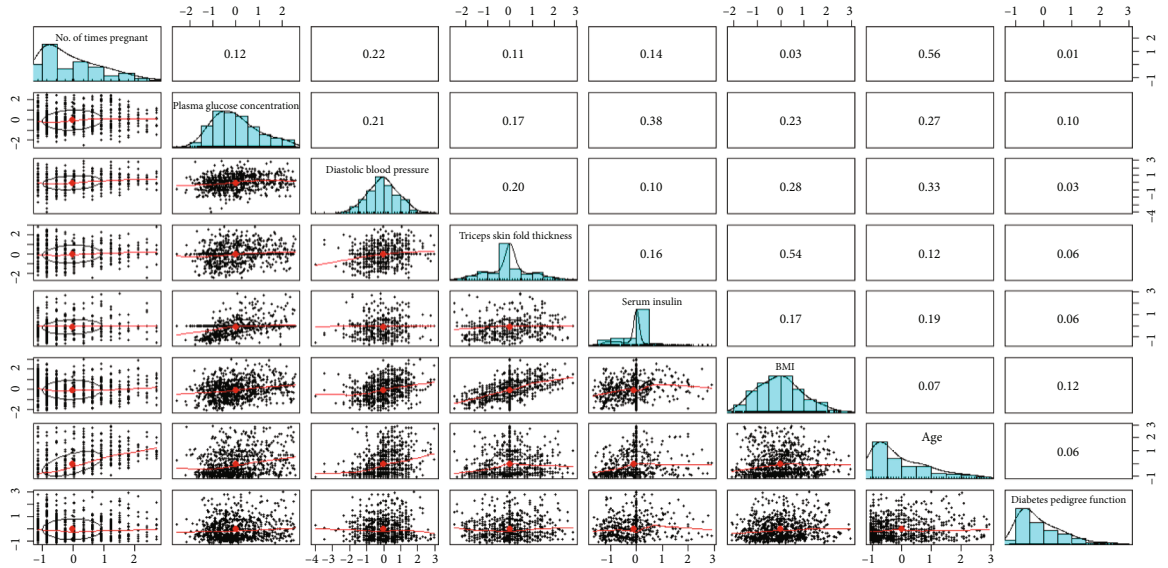

Figure 3: Scatter plots, bar charts, and correlation coefficients for the $M = 768$ dataset.

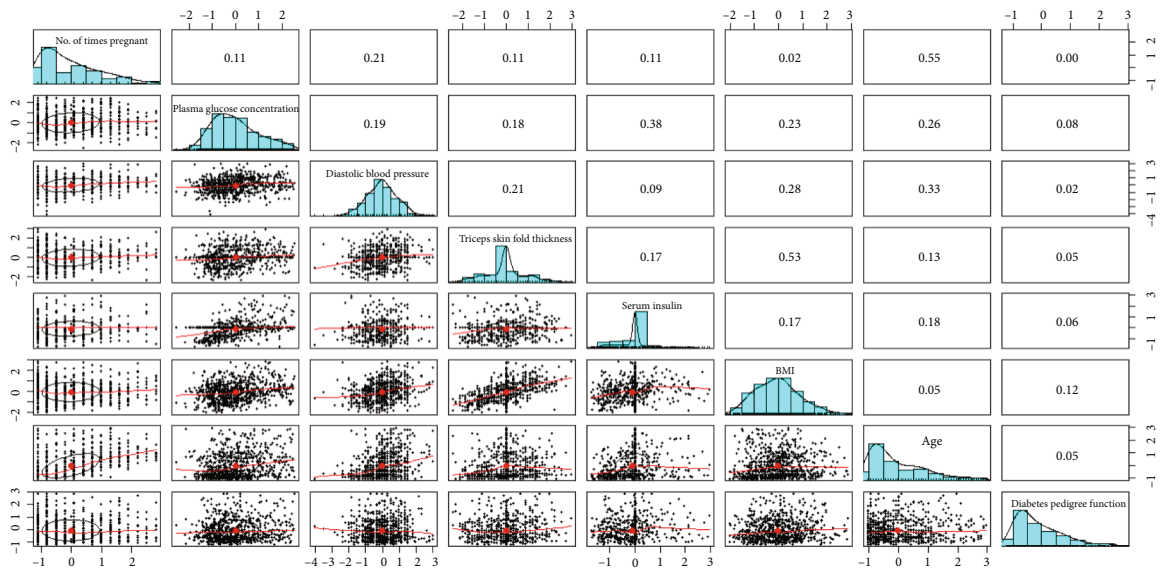

Figure 4: Scatter plots, bar charts, and correlation coefficients for the $M = 2000$ dataset.

and 6, green nodes represent the ancestor nodes, which have only child nodes; yellow nodes represent the intermediate nodes, which have both parent and child nodes; and orange nodes represent the child nodes, which have only parent nodes. Table 3 shows the maximum likelihoods for both structures.

(i) Similarities between structures 1 and 2: both structures exhibit six identical causal relationships: $X_1 \to X_7$, $X_7 \to X_3$, $X_4 \to X_6$, $X_5 \to X_2$, $X_2 \to X_6$, and $X_6 \to X_3$, with $X_1 \to X_7$ indicating that the number of times pregnant causes changes in age, $X_7 \to X_3$ indicating that age causes changes in diastolic blood pressure, $X_4 \to X_6$ indicating that triceps skin fold thickness causes changes in BMI, $X_5 \to X_2$ indicating that insulin causes changes in plasma glucose concentration, $X_2 \to X_6$ indicating that plasma glucose concentration causes changes in BMI, and $X_6 \to X_3$ indicating that BMI causes changes in diastolic blood pressure. There was an absence of causal relationships between diabetes pedigree function and all other variables in both structures

(ii) Differences between structures 1 and 2: structure 1 exhibits the causal relationship $X_6 \to X_7$, whereas structure 2 shows the causal relationships $X_7 \to X_2$ and $X_4 \to X_7$, with $X_6 \to X_7$ indicating that BMI causes changes in age, $X_7 \to X_2$ indicating that age causes changes in plasma glucose concentration, and $X_4 \to X_7$ indicating that triceps skin fold thickness causes changes in age

Table 1: Correlation coefficients and P values of variable pairs for the $M = 768$ dataset.

Variable pair	Correlation coefficient	P value
No. of times pregnant and age	0.56	0
No. of times pregnant and diastolic blood pressure	0.21	0
No. of times pregnant and insulin	0.14	0
No. of times pregnant and triceps skin fold thickness	0.11	0.003
No. of times pregnant and plasma glucose concentration	0.11	0.001
Plasma glucose concentration and insulin	0.38	0
Plasma glucose concentration and age	0.27	0
Plasma glucose concentration and BMI	0.23	0
Plasma glucose concentration and diastolic blood pressure	0.21	0
Plasma glucose concentration and triceps skin fold thickness	0.17	0
Plasma glucose concentration and diabetes pedigree function	0.10	0.005
Diastolic blood pressure and age	0.33	0
Diastolic blood pressure and BMI	0.28	0
Diastolic blood pressure and triceps skin fold thickness	0.20	0
Diastolic blood pressure and insulin	0.10	0.005
Triceps skin fold thickness and BMI	0.54	0
Triceps skin fold thickness and insulin	0.16	0
Triceps skin fold thickness and age	0.12	0.001
Insulin and age	0.19	0
Insulin and BMI	0.17	0
BMI and diabetes pedigree function	0.12	0.001

Table 2: Correlation coefficients and P values of variable pairs for the $M = 2000$ dataset.

Variable pair	Correlation coefficient	P value
No. of times pregnant and age	0.55	0
No. of times pregnant and diastolic blood pressure	0.21	0
No. of times pregnant and insulin	0.11	0
No. of times pregnant and triceps skin fold thickness	0.11	0
No. of times pregnant and plasma glucose concentration	0.11	0
Plasma glucose concentration and insulin	0.38	0
Plasma glucose concentration and age	0.26	0
Plasma glucose concentration and BMI	0.23	0
Plasma glucose concentration and diastolic blood pressure	0.19	0
Plasma glucose concentration and triceps skin fold thickness	0.18	0
Diastolic blood pressure and age	0.33	0
Diastolic blood pressure and BMI	0.28	0
Diastolic blood pressure and triceps skin fold thickness	0.21	0
Triceps skin fold thickness and BMI	0.53	0
Triceps skin fold thickness and insulin	0.17	0
Triceps skin fold thickness and age	0.13	0
Insulin and age	0.18	0
Insulin and BMI	0.17	0
BMI and diabetes pedigree function	0.12	0

Figure 3 shows that the correlation coefficient between BMI and age is 0.07, and the corresponding P value is 0.072. Therefore, the absence of a linear relationship between BMI and age can be deduced. Obviously, the causal function obtained by the regression method fails the significance test and has no statistical significance. On this basis, $X_6 \to X_7$ can be regarded as an erroneous causal relationship. In Figure 6, the erroneous causal edge $X_6 \to X_7$ was eliminated

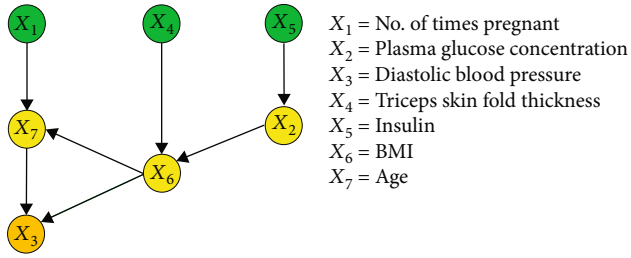

Figure 5: Causal structure for the $M = 768$ dataset (structure 1).

when causal discovery was performed with the $M = 2000$ dataset, but two additional causal edges $X_4 \rightarrow X_7$ and $X_7 \rightarrow X_2$ were discovered. As shown in Table 3, the maximum likelihood of structure 2 is higher than that of structure 1, which suggests that sample size influences the results of causal discovery. Although a larger sample size favours the elimination of erroneous causal edges and discovery of previously nonexistent causal edges, the increase in the number of discovered causal edges may also lead to an increase in the number of redundant edges. Figures 5 and 6 demonstrate that the causal structures were complex with significant numbers of redundant or erroneous edges, which necessitates the development of new causal discovery algorithms.

3.2.3. Results of IFCL-Based Causal Discovery. The purpose of this experiment was to compare the performance of the proposed method with the FCL-based causal discovery method and explore the optimal causal structure of diabetes risk factors. When the IFCL-based algorithm was adopted for causal discovery in the $M = 768$ and $M = 2000$ datasets, it was found that the results of causal discovery were closely associated with the adjustment threshold value. In the experiment, α values of $0.05 \leq \alpha \leq 0.18$ in intervals of 0.01 were adopted, whereas α values < 0.05 were not used owing to the generation of excessive redundant causal edges.

(i) Causal structures for the $M = 768$ dataset: Figure 7 shows the generated causal structure with five pairs of causal relationships ($X_1 \rightarrow X_7$, $X_3 \rightarrow X_6$, $X_4 \rightarrow X_6$, $X_5 \rightarrow X_2$, and $X_2 \rightarrow X_6$) when $\alpha = 0.05 - 0.06$ (structure 3). In structure 3, X_1, X_3, X_4, and X_5 are the ancestor nodes, X_2 is the intermediate node, and X_6 and X_7 are the child nodes. Compared with structure 1, $X_6 \rightarrow X_7$ (an erroneous edge) and $X_7 \rightarrow X_3$ are absent, and $X_6 \rightarrow X_3$ is reversed to form the $X_3 \rightarrow X_6$ relationship in structure 3. Figure 8 shows the generated causal structure with four pairs of causal relationships ($X_1 \rightarrow X_7$, $X_2 \rightarrow X_6$, $X_3 \rightarrow X_6$, and $X_4 \rightarrow X_6$) when $\alpha = 0.07 - 0.14$ (structure 4). In structure 4, X_1, X_2, X_3, and X_4 are the ancestor nodes, X_7 and X_6 are the child nodes, and there is no intermediate node. Compared with structure 3, the causal edge $X_5 \rightarrow X_2$ is absent in structure 4. Figure 9 shows the generated causal structure when $\alpha = 0.15$ (structure 5), which merely consists of two causal edges, $X_1 \rightarrow X_7$ and $X_4 \rightarrow X_6$. In structure 5, there are only the ancestor nodes (X_1 and X_4) and child nodes (X_7 and X_6). Further simplification did not occur in the causal structure when α was increased beyond 0.15

(ii) Causal structures for the $M = 2000$ dataset: Figure 10 shows the generated causal structure with six pairs of causal relationships ($X_1 \rightarrow X_7$, $X_4 \rightarrow X_7$, $X_3 \rightarrow X_6$, $X_4 \rightarrow X_6$, $X_5 \rightarrow X_2$, and $X_2 \rightarrow X_6$) when $\alpha = 0.05 - 0.06$ (structure 6). In structure 6, X_1, X_3, X_4, and X_5 are the ancestor nodes, X_2 is the intermediate node, and X_6 and X_7 are child nodes. Compared with structure 2, the causal edges $X_7 \rightarrow X_3$ and $X_7 \rightarrow X_2$ are absent, and $X_6 \rightarrow X_3$ is reversed to form the $X_3 \rightarrow X_6$ relationship in structure 6. Figure 11 shows the generated causal structure with five pairs of causal relationships ($X_1 \rightarrow X_7$, $X_4 \rightarrow X_7$, $X_4 \rightarrow X_6$, $X_3 \rightarrow X_6$, and $X_2 \rightarrow X_6$) when $\alpha = 0.07 - 0.15$ (structure 7). In structure 7, X_1, X_2, X_3, and X_4 are the ancestor nodes, X_7 and X_6 are the child nodes, and there is no intermediate node. Compared with structure 6, $X_5 \rightarrow X_2$ is absent from structure 7. When $\alpha = 0.16 - 0.17$, the algorithm could not find an optimal causal structure. Figure 12 shows the generated causal structure when $\alpha \geq 0.18$ (structure 8), which merely consists of two causal edges, $X_1 \rightarrow X_7$ and $X_4 \rightarrow X_7$. In structure 8, there are only the ancestor nodes (X_1 and X_4) and a child node (X_7). Additional changes did not occur in the causal structure when α was increased further

Table 4 shows the maximum likelihoods for structures 3–8. It can be seen that the maximum likelihood increased with increasing sample size.

The results presented above indicate that a larger sample size leads to a reduction in the number of erroneous causal relationships and the discovery of other potential causal relationships. During the causal discovery process, α must be incorporated to reduce the number of redundant and erroneous edges. When α was increased, the causal structures generated using the improved algorithm proposed in this study became increasingly simplified. In particular, when α was set to 0.05 or 0.06, causal structures with the fewest redundant edges and maximum information retention were obtained. Therefore, it can be deduced that the optimal adjustment threshold values for the discovery of causal relationships among the diabetes risk factors were 0.05 and 0.06.

4. Analysis and Discussion

As shown in Figures 7 and 10, a total of six causal relationships ($X_5 \rightarrow X_2$, $X_2 \rightarrow X_6$, $X_4 \rightarrow X_6$, $X_3 \rightarrow X_6$, $X_1 \rightarrow X_7$, and $X_4 \rightarrow X_7$), which are discussed in detail below, existed among the various diabetes risk factors.

(1) $X_5 \rightarrow X_2$, $X_2 \rightarrow X_6$: these causal relationships are well known among the general public. Insulin is the only hormone that lowers blood glucose levels in the human body. If insulin resistance occurs, abnormalities will arise in glucose uptake in the body, which

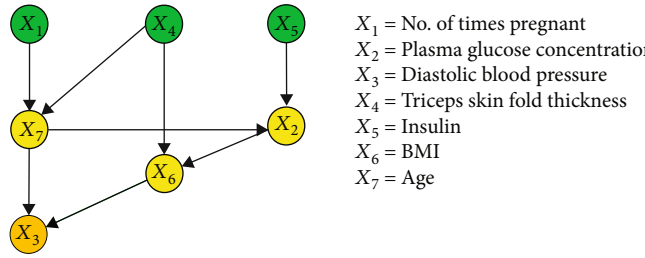

Figure 6: Causal structure for the $M = 2000$ dataset (structure 2).

Table 3: Maximum likelihoods of causal structures 1 and 2.

Causal structure	Maximum likelihood
1	-8.34
2	-8.17

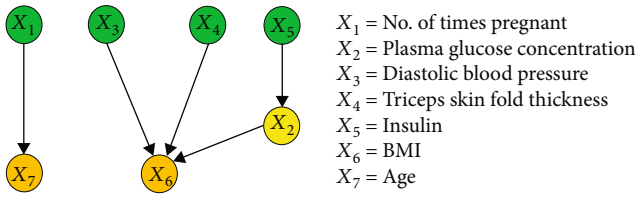

Figure 7: Causal structure for the $M = 768$ dataset ($\alpha = 0.05 - 0.06$) (structure 3).

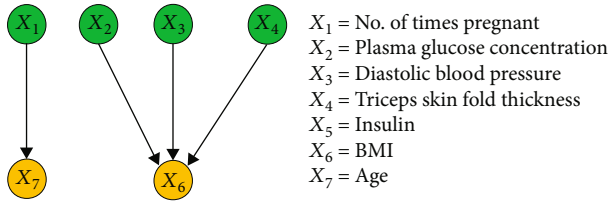

Figure 8: Causal structure for the $M = 768$ dataset ($\alpha = 0.07 - 0.14$) (structure 4).

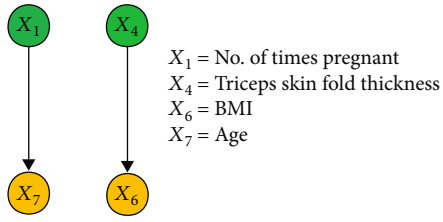

Figure 9: Causal structure for the $M = 768$ dataset ($\alpha = 0.15$) (structure 5).

will lead to increased plasma glucose concentration and result in a higher likelihood of diabetes onset. Additionally, $X_5 \rightarrow X_2$ and $X_2 \rightarrow X_6$ can be combined to form the causal relationship $X_5 \rightarrow X_2 \rightarrow X_6$. In a typical human body with normal insulin secretion, blood glucose metabolism will be at a standard level, which will lead to the maintenance of normal

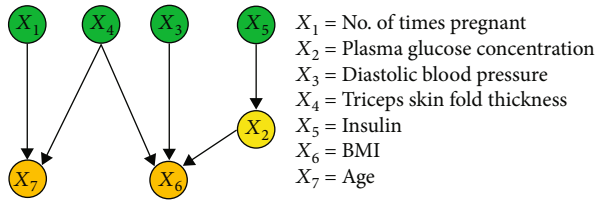

Figure 10: Causal structure for the $M = 2000$ dataset ($\alpha = 0.05 - 0.06$) (structure 6).

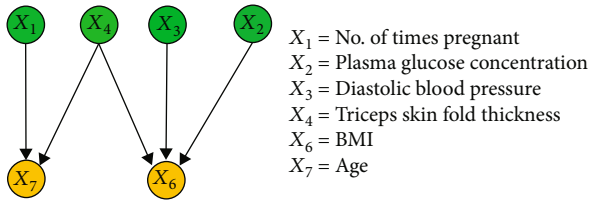

Figure 11: Causal structure for the $M = 2000$ dataset ($\alpha = 0.07 - 0.15$) (structure 7).

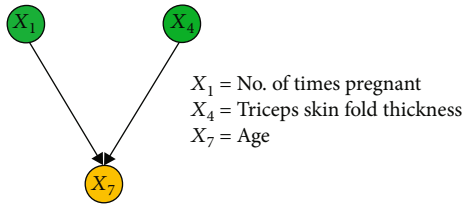

Figure 12: Causal structure for the $M = 2000$ dataset ($\alpha = 0.18$) (structure 8).

BMI. In contrast, in diabetic patients with insulin resistance, blood glucose cannot be effectively absorbed and utilised, which leads to decreased body weight and lower BMI. Therefore, the causal relationship $X_5 \rightarrow X_2 \rightarrow X_6$ also holds true

(2) $X_4 \rightarrow X_6$: the triceps skin fold thickness reflects body fat content, with a greater thickness indicating a higher body fat percentage and body weight, which leads to an increase in BMI and risk of diabetes onset. When diabetes causes emaciation in patients, triceps skin fold thickness and body weight are reduced, causing a decrease in BMI. Therefore, the causal relationship $X_4 \rightarrow X_6$ still holds true

(3) $X_3 \rightarrow X_6$: when causal discovery was performed in accordance with a previously reported method [33],

Table 4: Maximum likelihoods for causal structures 3–8.

Dataset	Causal structure	Maximum likelihood
$M = 768$	3 ($\alpha = 0.05 - 0.06$)	-8.18, -8.13
	4 ($\alpha = 0.07 - 0.14$)	-8.09, -8.03, -7.97, -7.91, -7.86, -7.79, -7.73, -7.67
	5 ($\alpha = 0.15$)	-7.63
$M = 2000$	6 ($\alpha = 0.05 - 0.06$)	-8.01, -7.96
	7 ($\alpha = 0.07 - 0.15$)	-7.92, -7.86, -7.80, -7.74, -7.68, -7.62, -7.56, -7.50, -7.44
	8 ($\alpha = 0.18$)	-7.26

the discovered relationship between factors 3 and 6 was $X_6 \rightarrow X_3$ (as shown in Figures 5 and 6), i.e., BMI influenced diastolic blood pressure. As people with higher body fat contents have higher BMIs and increased tendencies to develop hypertension, such a causal relationship is consistent with common medical knowledge and indicates that BMI is a trigger for hypertension. However, when causal discovery was performed using the modified method developed in this study, the reverse relationship ($X_3 \rightarrow X_6$) was discovered (as shown in Figures 7 and 10). This finding suggests the possible existence of a certain casual factor that changed under the influence of BMI and consequently influenced the risk of diabetes onset. Notably, certain diabetic patients suffer from concomitant hypertension and emaciation. Medical professionals generally believe that emaciation is caused by diabetes, but it may also be jointly influenced by diabetes and hypertension, resulting in changes in BMI. Therefore, $X_3 \rightarrow X_6$ may be a little-known relationship that exists in reality

(4) $X_1 \rightarrow X_7$: this causal relationship indicates that the number of times pregnant causes changes in age. In a previous study [35], it was reported that an increased number of pregnancies was associated with higher physiological age, i.e., cellular ageing may be accelerated, which in turn causes a higher probability of developing certain diseases. Therefore, the causal mechanism underlying $X_1 \rightarrow X_7$ may be as follows: an increased number of times pregnant causes accelerated ageing of pancreatic β cells, which leads to a higher tendency to develop insulin resistance and an increased diabetes risk

(5) $X_4 \rightarrow X_7$: this causal relationship indicates that the triceps skin fold thickness causes changes in age. As the triceps skin fold thickness reflects the nutritional status of an individual, the underlying causal mechanism for $X_4 \rightarrow X_7$ may be as follows: a triceps skin fold thickness that is less than the standard value indicates malnutrition, which affects physiological age and causes pancreatic β cell ageing, thereby causing insulin resistance and an increased diabetes risk. An excessively large triceps skin fold thickness indicates obesity, which signifies the presence of an excessive amount of glucose in the body. Consequently, the pancreatic β cells become overworked for long periods, which increases the tendency for ageing and functional damage in the pancreas, resulting in an increased risk of diabetes

In short, among the causal relationships identified through the IFCL-based causal discovery method proposed in this study, $X_5 \rightarrow X_2 \rightarrow X_6$ and $X_4 \rightarrow X_6$ are confirmed relationships, whereas $X_3 \rightarrow X_6$, $X_1 \rightarrow X_7$, and $X_4 \rightarrow X_7$ require further validation. These results suggest that the improved algorithm possesses huge potential for the discovery of causal relationships among diabetes risk factors and may be beneficial for further elucidation of causality among diabetes risk factors.

5. Conclusion

In the present study, we proposed an IFCL-based diabetes risk factor causal discovery algorithm that effectively resolves the issue of excessive redundant and erroneous edges in the causal structures generated by the FCL-based algorithm. Our experimental results demonstrate the efficacy of the proposed algorithm and provide a scientific basis for uncovering causal relationships among various diabetes risk factors. The next step in our research efforts will be the exploration of causality among the biochemical markers of diabetes and physiological indicators of body composition, with the objective of elucidating the causal relationships between the pathological and physiological factors of diabetes and enhancing diabetes prevention and treatment efforts.

Data Availability

The data used to support the findings of this study are included within the supplementary information files.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

We would like to thank Editage (http://www.editage.cn) for English language editing. This work is supported by the Competitive Allocation of Special Funds for Science and Technology Innovation Strategy in Guangdong Province of China (grant numbers 2018A06001 and 2019A1515011164).

Supplementary Materials

See Datasets S1–S2, Source Code S3, and Editorial Certificate S4 in the Supplementary Material for simulation experiments. *(Supplementary Materials)*

References

[1] J. Li, S. Wang, X. Han, G. Zhang, M. Zhao, and L. Ma, "Spatiotemporal trends and influence factors of global diabetes prevalence in recent years," *Social Science & Medicine*, vol. 256, p. 113062, 2020.

[2] E. Muzurović, Z. Stanković, Z. Kovačević, B. Š. Škrijelj, and D. P. Mikhailidis, "Inflammatory markers associated with diabetes mellitus—old and new players," *Current Pharmaceutical Design*, vol. 26, 2020.

[3] M. Fizelova, M. Miilunpohja, A. J. Kangas et al., "Associations of multiple lipoprotein and apolipoprotein measures with worsening of glycemia and incident type 2 diabetes in 6607 non-diabetic Finnish men," *Atherosclerosis*, vol. 240, no. 1, pp. 272–277, 2015.

[4] M. A. Lankinen, A. Stančáková, M. Uusitupa et al., "Plasma fatty acids as predictors of glycaemia and type 2 diabetes," *Diabetologia*, vol. 58, no. 11, pp. 2533–2544, 2015.

[5] S. Yazdanpanah, M. Rabiee, M. Tahriri et al., "Evaluation of glycated albumin (GA) and GA/HbA1c ratio for diagnosis of diabetes and glycemic control: a comprehensive review," *Critical Reviews in Clinical Laboratory Sciences*, vol. 54, no. 4, pp. 219–232, 2017.

[6] G. L. Huang, X. H. He, R. Pu, and Y. He, "Clinical value of adiponectin in combination with FGF-21 and A-FABP in early clinical diagnosis and risk prediction for type 2 diabetes," *Practical Combination of Traditional Chinese and Western Medicine*, vol. 17, no. 4, pp. 98-99, 2017.

[7] C. Bellia, M. Zaninotto, C. Cosma et al., "Clinical usefulness of glycated albumin in the diagnosis of diabetes: results from an Italian study," *Clinical Biochemistry*, vol. 54, pp. 68–72, 2018.

[8] Y. Tatsukawa, M. Misumi, Y. M. Kim et al., "Body composition and development of diabetes: a 15-year follow-up study in a Japanese population," *European Journal of Clinical Nutrition*, vol. 72, no. 3, pp. 374–380, 2018.

[9] H. Li, J. Lv, C. Yu et al., "The association between age at initiation of alcohol consumption and type 2 diabetes mellitus: a cohort study of 0.5 million persons in China," *American Journal of Epidemiology*, vol. 189, no. 12, pp. 1478–1491, 2020.

[10] C. Zhao, H. Wang, J. Zhang, and L. Feng, "Correlations of trace elements, glucose and body compositions in type 2 diabetes," *Health Research*, vol. 37, no. 5, pp. 600–605, 2008.

[11] T. Tillin, A. D. Hughes, Q. Wang et al., "Diabetes risk and amino acid profiles: cross-sectional and prospective analyses of ethnicity, amino acids and diabetes in a South Asian and European cohort from the SABRE (Southall And Brent REvisited) study," *Diabetologia*, vol. 58, no. 5, pp. 968–979, 2015.

[12] Y. Cui and Z. P. Feng, "Relationship between serum 25-(OH) D level and body fat distribution in patients with type 2 diabetes," *Chinese Journal of Osteoporosis*, vol. 22, no. 1, pp. 56–60, 2016.

[13] T. Huang, K. Glass, O. A. Zeleznik et al., "A network analysis of biomarkers for type 2 diabetes," *Diabetes*, vol. 68, no. 2, pp. 281–290, 2019.

[14] N. N. Zhu, Y. X. Liu, S. S. Wang, R. Geng, Y. Liu, and D. Li, "Relationship between level of glycemic control and body composition in patients with type 2 diabetes," *Chinese Journal of Diabetes*, vol. 27, no. 3, pp. 42–45, 2019.

[15] Z. Liu, Y. Zhang, S. Graham et al., "Causal relationships between NAFLD, T2D and obesity have implications for disease subphenotyping," *Journal of Hepatology*, vol. 73, no. 2, pp. 263–276, 2020.

[16] K. Chien, T. Cai, H. Hsu et al., "A prediction model for type 2 diabetes risk among Chinese people," *Diabetologia*, vol. 52, no. 3, pp. 443–450, 2009.

[17] Y. Y. Li, R. Li, and S. N. Zhang, "Evaluation on effect of screening method for undiagnosed diabetes," *Chinese Journal of Public Health*, vol. 6, pp. 687–689, 2006.

[18] Z. Xin, J. Yuan, L. Hua et al., "A simple tool detected diabetes and prediabetes in rural Chinese," *Journal of Clinical Epidemiology*, vol. 63, no. 9, pp. 1030–1035, 2010.

[19] D. Noble, R. Mathur, T. Dent, C. Meads, and T. Greenhalgh, "Risk models and scores for type 2 diabetes: systematic review," *BMJ*, vol. 343, no. nov28 1, 2011.

[20] A. Abbasi, E. Corpeleijn, L. M. Peelen et al., "External validation of the KORA S4/F4 prediction models for the risk of developing type 2 diabetes in older adults: the PREVEND study," *European Journal of Epidemiology*, vol. 27, no. 1, pp. 47–52, 2012.

[21] S. K. Tanamas, D. J. Magliano, B. Balkau et al., "The performance of diabetes risk prediction models in new populations: the role of ethnicity of the development cohort," *Acta Diabetologica*, vol. 52, no. 1, pp. 91–101, 2015.

[22] A. Rahman, K. Nesha, M. Akter, and S. Uddin, "Application of artificial neural network and binary logistic regression in detection of diabetes status," *Science Journal of Public Health*, vol. 1, no. 1, pp. 39–43, 2013.

[23] Y. Isler, A. Narin, M. Ozer, and M. Perc, "Multi-stage classification of congestive heart failure based on short-term heart rate variability," *Chaos, Solitons & Fractals*, vol. 118, pp. 145–151, 2019.

[24] X. Wang and D. F. Chen, "Application of support vector machine on predictive model of type 2 diabetes," *Chinese Journal of Chronic Disease Prevention and Control*, vol. 18, no. 6, pp. 560–562, 2010.

[25] J. Song, X. S. Wu, J. Zhang, Y. Y. Zhang, and X. ChenApplication of three statistical models in the prediction of diabetes risk," *China Health Statistics*, vol. 34, no. 2, pp. 312–314, 2017.

[26] Y. Chen, H. J. Zong, and W. Li, "A research on risk factors and risk prediction models of type 2 diabetes mellitus," *Journal of Kunming University of Science and Technology (Natural Science Edition)*, vol. 43, no. 2, pp. 60–64, 2018.

[27] O. Erkaymaz, M. Ozer, and M. Perc, "Performance of small-world feedforward neural networks for the diagnosis of diabetes," *Applied Mathematics and Computation*, vol. 311, pp. 22–28, 2017.

[28] O. Geman, I. Chiuchisan, and R. Toderean, "Application of adaptive neuro-fuzzy inference system for diabetes classification and prediction," in *2017 E-Health and Bioengineering Conference (EHB)*, pp. 639–642, Sinaia, Romania, 2017.

[29] L. S. Liu, Q. Li, F. Yang, Z. Z. Zheng, X. J. Lin, and Q. Wu, "Classification method of diabetes based on integration of characteristic classifier," *Chinese Journal of Traditional Chinese Medicine*, vol. 31, no. 1, pp. 80–83, 2016.

[30] B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, "Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction," *Artificial Intelligence in Medicine*, vol. 85, pp. 43–49, 2018.

[31] Y.-T. Wu, C.-J. Zhang, B. W. Mol et al., "Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning," *The Journal of Clinical Endocrinology & Metabolism*, vol. 106, no. 16, 2021.

[32] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, UK, 2009.

[33] R. Cai, J. Qiao, Z. Zhang, and Z. Hao, "SELF: structural equational embedded likelihood framework for causal discovery," in *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 1787–1794, New Orleans, LA, USA, 2018.

[34] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*, MIT Press, Cambridge, MA, 2000.

[35] C. P. Ryan, M. G. Hayes, N. R. Lee et al., "Reproduction predicts shorter telomeres and epigenetic age acceleration among young adult women," *Scientific Reports*, vol. 8, no. 1, 2018.