

Research on the Correlation Between Multiple Risk Factors and Diabetes Mellitus

Yuzhen Li^{1,*} and Haoze Song²

¹College of Arts and Sciences, The Ohio State University, Columbus, 43210, U.S.A.

²School of Science, China University of Mining and Technology-Beijing, Beijing, 100083, China

Keywords: Diabetes Mellitus, Risk Factors, Binary Probit Model.

Abstract: Preventing diabetes is one of the main areas of current study. Numerous studies have revealed a strong correlation between smoking, age, obesity, and the prevalence of diabetes; nonetheless, a great deal more study has to be done in this area. Therefore, in this paper, the correlation between multiple risk factors and diabetes was investigated by collecting data on the physical status of patients at Sylhet Diabetes Hospital and performing a binary Probit analysis on the data. Polyuria, polydipsia, polyphagia, genital thrush, irritability, and partial paresis were found to be strong risk factors for developing diabetes, while sudden weight loss, weakness, visual blurring, delayed healing, muscle stiffness, alopecia, and obesity were found to be less influential in causing diabetes. The study enables greater attention to be paid to the risk factors associated with diabetes, facilitates early detection of diabetes and early initiation of treatment to improve patient prognosis and survival, and points to a clear direction for future related research.

1 INTRODUCTION

The history of diabetes can be dated back to several millennia ago, especially in ancient Egypt, Greece, and Asia. Under biological settings, the kidney maintains fluid and electrolyte homeostasis by modifying the amount and chemical makeup of urine in accordance with the body's requirements. A complicated and varied clinical disease known as diabetes affects the distribution of water and is typified by persistent diuresis, which produces huge amounts of diluted urine (Valenti and Tamma, 2016).

In recent years, the global incidence of diabetes has climbed dramatically. Diabetes affects 529 million people globally, a figure that is anticipated to rise to 1.3 billion by 2050 (Liu, 2023). Drawing on information from the "Global Burden of Disease 2021" research, the freshest and most extensive assessments to date indicate that diabetes now affects 6.1% of the global population. This positions diabetes as a leading contributor to mortality and disability worldwide, ranking it among the top ten causes. The findings underscore the significant and growing impact of diabetes on global health, necessitating urgent and sustained efforts to address this escalating public health challenge (Liu, 2023). By 2050, it is anticipated that the prevalence of obesity in North

Africa and the Middle East will skyrocket from 9.3 percent to a staggering 16.8 percent. In a similar vein, projections indicate that the occurrence rate in Latin America and the Caribbean is expected to rise significantly, reaching 11.3 percent (Liu, 2023). Research conducted in the United States and Europe has demonstrated a 3-5% annual increase in the incidence of type 1 diabetes (T1D) throughout time (Maffi and Secchi, 2017). Diabetes is notably common among individuals aged 65 and above, affecting over 20% of this demographic worldwide. The occurrence is particularly pronounced in those between 75 and 79 years old (Liu, 2023). To prevent a rise in the incidence of diabetes, it is essential to implement measures aimed at assessing the degree of correlation between diabetes and various risk factors. Comprehensive studies should be conducted to thoroughly investigate how these risk variables interact with the development and progression of diabetes, thereby enabling the formulation of effective prevention strategies.

The development of diabetes mellitus involves multiple pathogenic processes, and the etiologic classification is very complicated, making it difficult to diagnose individuals with risk factors (Li et al., 2024). National and international scholars have found smoking, age, and obesity to be factored in the onset

and development of diabetes (Li et al., 2019; Gong et al., 2017 & Qiu et al., 2016). In addition, God et al. found that other lifestyle factors also influence the incidence of diabetes (Gode et al., 2024). However, the risk factors of diabetes mellitus in this literature failed to be adequately studied. In addition, the statistically obtained data are not systematic and complete. As a result, this study concentrates on examining a set of 16 distinct variables, which include Gender, Age, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, and Obesity. The main goal of this research is to discern and utilize the most appropriate statistical model to evaluate the degree of association between the identified risk factors and the prevalence of diabetes. Through this analysis, the study aspires to offer a thorough comprehension of the individual contribution of each factor to the probability of developing diabetes. This understanding is critical for the formulation of more precise and impactful prevention and intervention strategies. By elucidating the specific roles these risk factors play, the research aims to enhance the effectiveness of public health initiatives and clinical practices aimed at mitigating the incidence of diabetes.

In a similar direction, Xue et al. used a variant of the Cox proportional risk model (Xue et al., 2022), but due to data limitations, they did not include individuals aged 35-45 years, whereas diabetes is also highly prevalent in the 35-45 year age group, so this may have affected the generalizability of the results. In addition, the study did not use biomarkers (e.g., blood glucose levels or HbA1c) to confirm the diagnosis of diabetes but relied on self-reporting and medication use, which may have affected the accuracy of the results. Ampeir et al. used a logistic regression model (Ampeire, Kawugeze and Mulogo, 2023), which provided either an Odds Ratio (OR) or an Adjusted Odds Ratio (AOR), which allows researchers to quantify the extent to which each predictor variable affects prediabetes risk. Nevertheless, it should be noted that the model presupposes a linear association between the dependent variable, which in this case is prediabetes, and the set of independent variables. This assumption implies that changes in the independent variables are directly proportional to changes in the likelihood of prediabetes, simplifying the complexity of potential nonlinear interactions that might exist in reality. If the actual relationship is nonlinear, the model may not accurately capture this relationship, leading to

inaccurate predictions and interpretations. Narjes Hazar et al. constructed a Dersimonian and Laird random effects model (Hazar et al., 2024). The model can handle the heterogeneity that exists between studies. However, the estimation of heterogeneity is sensitive to the distribution of the data and the size of the study. If the sample size of some studies is extremely large or small, it may have an unbalanced effect on the overall heterogeneity estimate.

By emphasizing the most important risk variables that ought to be the focus of intervention, the results of this study are anticipated to guide the practice of medicine, influence public health efforts, and offer novel perspectives on early detection of diabetes tactics. Furthermore, the discovery of neglected variables can provide avenues for further investigation, which would ultimately lead to a more thorough comprehension of the genesis of diabetes.

2 METHODOLOGY

2.1 Data Sources

The dataset utilized for this research has been sourced from the repository available on Kaggle's platform. The data for this study was gathered through the administration of direct questionnaires to patients receiving care at Sylhet Diabetes Hospital, located in Sylhet, Bangladesh. The collected information was subsequently reviewed and authorized by a medical professional to ensure its accuracy and reliability.

2.2 Variable Selection

This study utilizes data from a cohort comprising 520 individuals, including both diabetic and non-diabetic patients. Within this population, there are 328 males and 192 females. The age distribution of these subjects spans from 16 to 90 years. The data contains 16 variables (Gender, Age, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, Obesity).

Table 1: Symbols and numerical representations of the sixteen variables.

Elements	Ideogram	Number1	Diabetes1
Gender	x ₁	520	320
Age	x ₂	520	320
Polyuria	x ₃	258	243
Polydipsia	x ₄	233	225
Sudden weight loss	x ₅	217	188
Weakness	x ₆	305	218
Polyphagia	x ₇	237	189
Genital thrush	x ₈	116	83
Visual blurring	x ₉	233	175
Itching	x ₁₀	253	154
Irritability	x ₁₁	126	110
Delayed healing	x ₁₂	239	153
Partial paresis	x ₁₃	224	192
Muscle stiffness	x ₁₄	195	135
Alopecia	x ₁₅	179	78
Obesity	x ₁₆	88	61

*Number 1: The prevalence of individuals afflicted by the disease.

**Diabetes1: The prevalence of individuals diagnosed with diabetes mellitus who are experiencing the condition.

Table 1 presents the statistics of the general population alongside the count of individuals diagnosed with diabetes mellitus. The data sample is 520 individuals, 320 of whom have diabetes.

Table 2 provides a detailed breakdown of the population, categorizing individuals by gender-females and males-as well as by diabetic status across various age brackets. The majority of the dataset consists of individuals aged between 31 and 60 years.

2.3 Methodology Introduction

This study employs a binary Probit regression analysis model to investigate the incidence of

diabetes, designated as the dependent variable (Y), in relation to 16 independent variables (X). For the dependent variable, a value of 0 represents the absence of diabetes, while a value of 1 signifies its presence. Subsequently, the influence of each of the 16 independent variables on the likelihood of having diabetes is examined. This analysis is conducted using SPSSAU software to facilitate the prediction of early-stage diabetes risk. By comprehensively assessing the relationship between these factors and diabetes, this research aims to provide valuable insights for early diagnosis and intervention strategies.

2.4 Model Testing

The accuracy of the model's predictions serves as an indicator of how well the model fits the data. According to the data presented in Table 3, it is clear that the research model has achieved an overall prediction accuracy of 93.27%. This high level of accuracy indicates that the model's performance is within acceptable parameters. When we delve deeper into the specifics, we find that the model exhibits a prediction accuracy of 91.50% when the actual value is 0. On the other hand, the accuracy improves to 94.38% when the actual value is 1. These figures collectively highlight the model's strong performance across various conditions, thereby validating its reliability and effectiveness for predictive purposes. The consistency in accuracy across different actual values underscores the robustness of the model, making it a dependable tool for predictive analytics. Such performance metrics not only demonstrate the model's capability to generalize well across different scenarios but also affirm its potential application in real-world predictive tasks.

Table 2: Gender distribution by age group.

Age	[11-20]	[21-30]	[31-40]	[41-50]	[51-60]	[61-70]	[71-80]	[81-90]
Number2	1	44	120	145	127	66	10	4
Female	0	13	59	55	39	22	0	2
Male	1	31	61	90	88	44	10	2
Diabetes2	0	34	69	94	80	36	7	0

*Number 2: The population distribution across different age brackets.

**Diabetes 2: The number of individuals across various age categories diagnosed with diabetes mellitus.

Table 3: Overview of Prediction Accuracy in Binary Probit Regression.

	Forecast value		Prediction accuracy	Prediction error rate
	0	1		
Actual value	0	183	91.50%	8.50%
	1	18	94.38%	5.63%
Summary			93.27%	6.73%

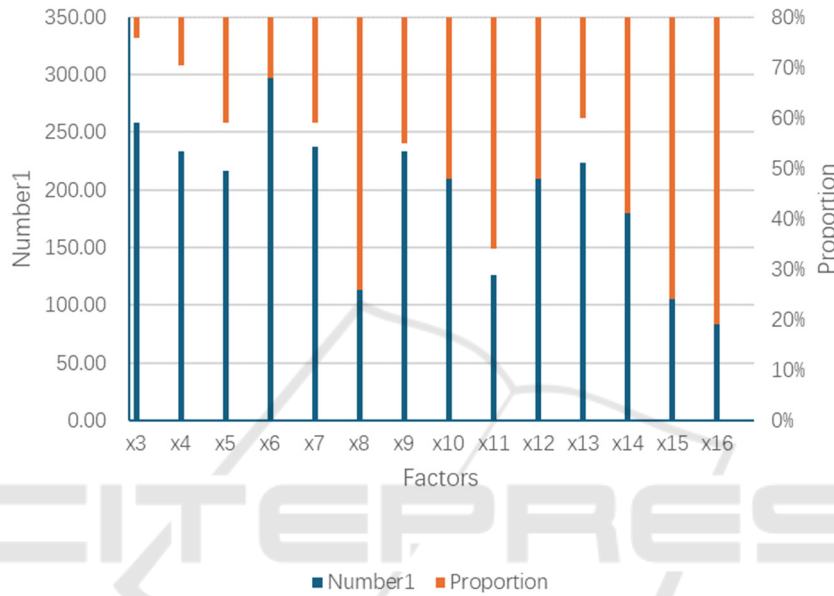


Figure 1: Factors of the prevalence of diabetes in the population.

3 RESULTS AND DISCUSSION

3.1 Descriptive Analysis

Figure 1 presents the percentage of individuals diagnosed with diabetes and enumerates several contributing factors associated with the disease. These contributing factors include Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, and Obesity. The dataset comprises a total of 520 samples. To determine the probability of each potential risk factor leading to diabetes, the percentage of diabetic samples identified and categorized based on various influencing variables was meticulously calculated. This analysis provides a comprehensive understanding of the

correlation between each risk factor and the likelihood of developing diabetes, thereby offering valuable insights into the etiology of the disease.

It is evident from Figure 1 above that Polyuria makes up the largest percentage, reaching up to 76%. With 19 percent, obesity had the lowest share. The data in the figure indicates that polyuria is the most prevalent trigger, which raises the biggest concealed risk of diabetes.

The analysis presented in the comparison chart reveals a pronounced disparity between the proportion of individuals diagnosed with diabetes and those without the condition. This discrepancy is particularly striking, with the most substantial difference being as high as 68.5%. The data underscore the significant variation in diabetes prevalence, highlighting the need for further investigation into the factors contributing to this considerable gap (Figure 2).

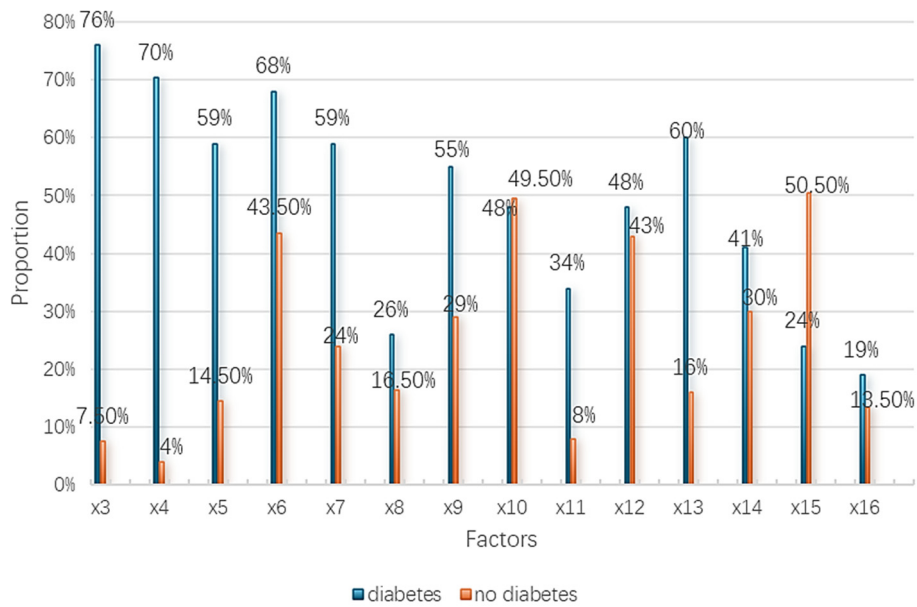


Figure 2: Comparative analysis of diabetic and non-diabetic individuals across various factors.

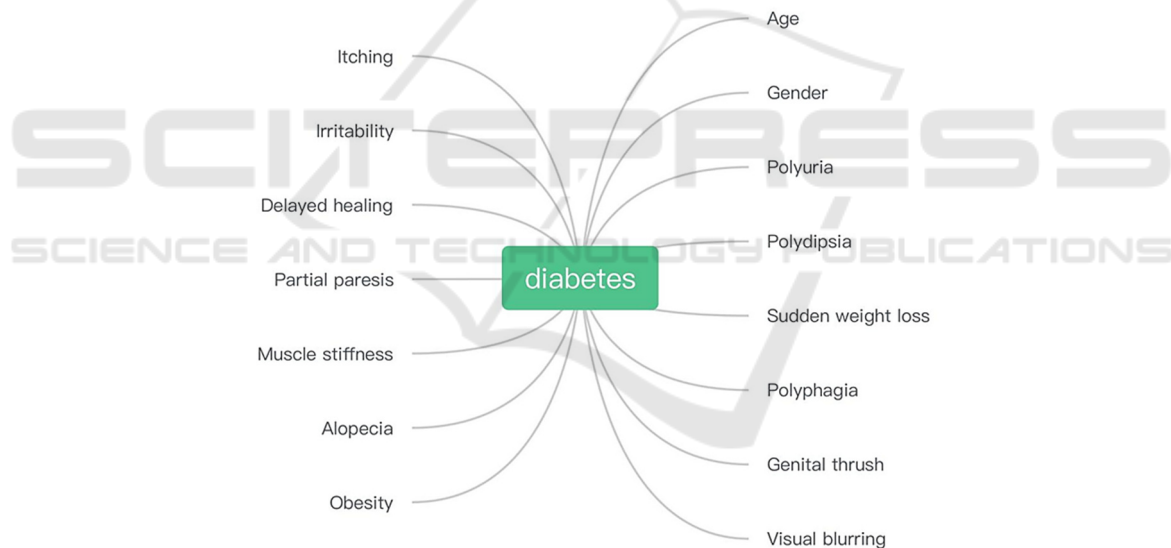


Figure 3: Variable-related schematic.

As illustrated in Figure 3, the study incorporated a range of variables potentially associated with diabetes into the analytical model. These variables encompass a comprehensive list of factors such as Gender, Age, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, and Obesity. By conducting a series of computations, this research ultimately derives the concluding linear regression formula:

$$Probit(p) = -4.079 - 0.028x_1 - 2.349x_2 + \dots - 0.119x_{15} \quad (1)$$

In this context, p denotes the likelihood that the variable for diabetes equals 1.

3.2 Binary Probit Model Results

Table 4 clearly illustrates that the model assesses whether the p-value exceeds the threshold of 0.05. This evaluation enables the study to delve deeper into identifying potential risk factors associated with diabetes mellitus. Through this detailed analysis, we

can infer connections that might otherwise remain undetected, thereby contributing valuable insights to the ongoing research on diabetes risk factors. On the one hand, diabetes is unaffected by the p values of abrupt weight loss, weakness, blurred vision, delayed healing, muscular stiffness, baldness, and obesity, all of which are bigger than 0.05. Conversely, the p values for partial paresis, age, and polyphagia fall within the range of greater than 0.01 but less than 0.05. This statistical outcome implies that these variables exert a notable influence on the development or progression of diabetes. Specifically, the p values associated with age, polyphagia, and partial paresis, while exceeding 0.01, still remain below 0.05, thereby underscoring their significant contribution to the condition.

This statistical evidence highlights the meaningful impact of these factors, suggesting that they play an important role in the pathophysiology of diabetes. Such findings are critical for understanding the multifaceted nature of diabetes and underscore the

importance of considering these variables in both clinical and research settings. However, the p values for gender, polyuria, polydipsia, genital thrush, itching, and irritability are less than 0.01, indicating that these factors have an especially significant association with diabetes. In terms of influence relationships, the regression coefficient and marginal effect values reveal the specific factors impacting diabetes. For each unit increase in Polyuria, Polydipsia, Polyphagia, Genital Thrush, Irritability, and Partial Paresis, the incidence of diabetes is projected to rise by 852.30%, 827.08%, 223.65%, 359.85%, 428.05%, and 230.99%, respectively. Given that each of these statistics is significant when compared to one another, it is evident that these elements play a crucial role in influencing the likelihood of diabetes onset. This suggests that the data have some research and a value of reference.

Table 4: Results overview from the binary Probit regression analysis.

Ideogram	regression coefficient	SE	z-value	p-value	95% CI	marginal effect
x ₁	-0.028	0.013	-2.082	0.037	-0.054 ~ -0.002	-0.003
x ₂	-2.349	0.301	-7.791	0.000	-2.940 ~ -1.758	-0.217
x ₃	2.380	0.353	6.748	0.000	1.689 ~ 3.072	0.220
x ₄	2.792	0.426	6.547	0.000	1.956 ~ 3.628	0.258
x ₅	0.188	0.289	0.651	0.515	-0.378 ~ 0.754	0.017
x ₆	0.447	0.282	1.584	0.113	-0.106 ~ 1.001	0.041
x ₇	0.630	0.287	2.195	0.028	0.067 ~ 1.192	0.058
x ₈	1.049	0.308	3.410	0.001	0.446 ~ 1.653	0.097
x ₉	0.479	0.351	1.366	0.172	-0.208 ~ 1.167	0.044
x ₁₀	-1.521	0.349	-4.354	0.000	-2.205 ~ -0.836	-0.141
x ₁₁	1.270	0.319	3.984	0.000	0.645 ~ 1.895	0.117
x ₁₂	-0.226	0.296	-0.766	0.444	-0.806 ~ 0.353	-0.021
x ₁₃	0.638	0.283	2.258	0.024	0.084 ~ 1.193	0.059
x ₁₄	-0.346	0.301	-1.150	0.250	-0.935 ~ 0.244	-0.032
x ₁₅	0.178	0.319	0.559	0.576	-0.447 ~ 0.804	0.016
x ₁₆	-0.119	0.307	-0.389	0.697	-0.721 ~ 0.482	-0.011
Constant	-4.079	0.915	-4.456	0.000	-5.874 ~ -2.285	-

3.3 Comparison with Existing Literature

In several earlier studies, researchers frequently employed a narrow examination of a single component, such as genetic inheritance, which was restricted to established risk factors for diabetes development. Additionally, several scholars suggest that the increased incidence of diabetes may be linked to the adoption of a Western lifestyle. This assertion is supported by migration studies, which highlight the transition from conventional agricultural practices to contemporary American living as a contributing factor (Wu et al., 2022). On the other hand, the factors included in this study are more thorough, which can help prevent errors that arise from failing to account for a single variable. It can also expand the ideas of forthcoming studies on diabetes, assisting medical professionals in identifying additional treatment avenues and early diabetes detection for prompt treatment.

4 CONCLUSION

Based on these experimental results, the main conclusion can be summarized as follows. 16 variables may lead to diabetes, in this study, a binary Probit regression analysis model is used and indicates that the most significant risk factors identified by these analyses and the data in the table are polyuria, polydipsia, polyphagia, genital thrush, irritability, and partial paresis. Sudden weight loss, weakness, visual blurring, delayed healing, muscle stiffness, alopecia, and obesity have a weak impact on causing diabetes.

However, it should be noted that this study has examined only some people from 11 to 90, which means that this conclusion may not be used to study the risk factors for children who are less than 11 years old and infants. And it is undeniable that this model may have mistakes in addition to the components because of the small amount of data, and that the sample did not include all ethnicities, which might have resulted in differences that could have also affected the accuracy of the findings. Notwithstanding its limitations, this research reported here would seem to indicate some valuable aspects. Firstly, the analysis indicates that Polyuria is the most common trigger and poses the highest hidden risk for diabetes, as demonstrated through a graphical method used to visually compare the proportions of each factor among individuals. Further research is required to prevent Polyuria from developing into diabetes,

and more drugs are being developed to slow the disease's progression to provide a complete picture of the diabetes treatment. Secondly, researchers create a representation of the experiment and make the results easier to comprehend and intuitive by comparing relative data, which includes those with diabetes and those without the disease. These tables enable people to pay more attention and emphasize the factors related to diabetes. The results of this study will guide future relevant research since it is necessary to determine whether these characteristics are linked to the development of diabetes. By identifying new causal components in addition to previously known causes, the prognosis and survival of patients can be improved through early detection of diabetes and prompt initiation of therapy.

AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

REFERENCES

- Valenti G and Tamma G 2016 History of Diabetes Insipidus. *Giornale italiano di nefrologia (Aldo Moro: Biotechnologies and Biopharmaceutics University of Bari)*.
- Xia Liu 2023 By 2050 there will be 1.3 billion people with diabetes worldwide. *Science and Technology Daily*. 3465.
- Maffi P and Secchi A 2017 The Burden of Diabetes: Emerging Data. *Management of Diabetic Retinopathy*. **60** 1-5.
- Li T, Sun Y Y, Li X L and Dong H 2024 Auxiliary diagnosis of diabetes mellitus based on machine learning classification algorithm. *Computer Knowledge and Technology* **20** 27-29.
- Li H Q, Liu L J and Xu Y C 2019 Effects of onset age on islet function and related indicators of newly diagnosed type 2 diabetes mellitus. *Public Health and Preventive Medicine* **30** 134-137.
- Gong H Y, Liu X F, He Y, Wang J Y and Deng Y 2017 The interaction of obesity and smoking on diabetes mellitus. *Chronic Disease Prevention and Control in China*. **25** 592-594.
- Qiu J Y Z, Hou X H and Jia W P 2016 Research progress on the correlation between smoking and diabetes mellitus. *Journal of Shanghai Jiao Tong University (Medical Science Edition)* **36** 110-114.
- Gode Y, Patond S, Wankhade V, Ghodki S, Jadhav D, Dhawade M R and Wankhade Y 2024 Study of Impact of Lifestyle Modification on Diabetes and Prediabetes

- in an Urban Population. *E3S Web of Conferences* **491** 3002.
- Xue L, Wang H, He Y, Sui M, Li H, Mei L and Ying X 2022 Incidence and risk factors of diabetes mellitus in the Chinese population: a dynamic cohort study. *BMJ Open* **12**.
- Ampeire I P, Kawugezi P C and Mulogo E M 2023 Prevalence of prediabetes and associated factors among community members in rural Isingiro district. *BMC Public Health* **23** 958.
- Hazar N, Jokar M, Namavari N, Hosseini S and Rahmanian V 2024 An updated systematic review and Meta-analysis of the prevalence of type 2 diabetes in Iran, 1996-2023. *Frontiers in Public Health*. **12**.
- Wu Y L, Yu Y W, Zhou J, Wang Y Y, Yu L S, Zhang J and Liu T 2022 Prospective study on the relationship between healthy lifestyle and diabetes in prediabetic population. *Modern Preventive Medicine* **49** 1350-1355.

