Measurement: Sensors 24 (2022) 100543

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

Identifying risk factors associated with type 2 diabetes based on data analysis

Waleed Noori Hussein [a,*], Zainab Muzahim Mohammed [b], Amani Naama Mohammed [b]

[a] Physiology Department, AL-Zahraa College of Medicine, University of Basrah, Basrah, Iraq
[b] Biochemistry Department, AL-Zahraa College of Medicine, University of Basrah, Basrah, Iraq

ARTICLE INFO

Keywords:
Data analysis
Diabetes
Computer system
Risk factors

ABSTRACT

This paper aims to identify risk factors and elements associated with type 2 diabetes. Both quantitative and qualitative approaches were used to collect data. Risk factors for type 2 diabetes are also presented from the systematic literature review to determine the variables and which variable seems to have the greatest impact on type 2 diabetes. Factor analysis of the dataset is used to provide an efficient result to predict and evaluate type 2 diabetes. This paper focused on increasing the accuracy of understanding type 2 diabetes based on data analysis. The results showed that Body Mass Index (BMI) has a strong influence on hemoglobin (A1C) with R-squared (R2) of 78%, and 60% with triglyceride (TG). The outcome of this paper will be a prediction model using a PLC-regression outer model analysis.

1. Introduction

Diabetes Mellitus (DM), generally known as diabetes, is a chronic metabolic condition characterized by elevated blood glucose levels [1]. It is one of the top 10 global killers, causing 4.2 million lives in 2019 [2], with type2 (T2DM) being the commonest, which is brought up by improper body response to insulin [3]. Diabetes is believed to be influenced by a variety of factors, these include inactivity, a poor diet, tobacco use, and alcohol intake [4]. In addition, obesity is a major contributor to nearly 55% of T2DM, accounting for the noticed increment in T2DM among teens in the 60s and 90s [5]. The prevention of type 2 diabetes mellitus (T2DM) is listed as a targeted concern by the World Health Organization as well as the United Nations in the 2018 Berlin Statement [1]. TD2M is now estimated to impact up to 463 million people worldwide including 1 in 11 adults [23]. Since the 1990s, prediabetes and type 2 diabetes have become more common among children. These conditions are linked to a high-fat diet, sedentarism, obesity, and disorders of the liver [2]. HbA1c has also been acknowledged as a measure to evaluate secondary vascular problems brought on by metabolic derailments in susceptible people. However, population-specific cut-offs may be required due to ethnic disparities in the sensitivity and specificity of HbA1c measurements [3]. Moreover, machine learning and data mining, are two examples of artificial intelligence (AI) technologies that are progressively integrating with medical science and becoming significant drivers of medical advancement [8]. Studies on type 2 diabetes conducted on computers have looked at or presented models that specify and describe elements that must be considered to assist in controlling the disease. A machine learning model was employed by University of Toronto researchers to examine health information on 2.1 million residents of Ontario that was gathered between 2006 and 2016. They discovered that they could use the algorithm to precisely forecast how many patients would get type 2 diabetes over five years. The machine learning algorithm could also determine if a person had a high or low probability of contracting the disease by analyzing several risk variables [9].

This paper aims to investigate and determine the risk factors and elements which can develop or prevent type 2 diabetes. In addition, this study aims to improve the accuracy of understanding for treating diabetes conditions. To that end, the very next section of this paper will review previous related studies. The methodology is discussed in the third section, discussion of the results is in the fourth section. The fifth section is the conclusion.

2. Review of previous related studies

Table 1 presents a summary of the past related studies on tools and methods used for analyzing diabetes.

The majority of previous research used SPSS and machine learning to

* Corresponding author.
E-mail address: waleed.hussein@uobasrah.edu.iq (W.N. Hussein).

https://doi.org/10.1016/j.measen.2022.100543
Received 16 September 2022; Received in revised form 28 September 2022; Accepted 24 October 2022
Available online 29 October 2022

Table 1
Summary of previous related studies.

Tool and method used for Analysis	Description	results	Author
SPSS (*t*-test and ANOVA)	To assess the effectiveness and feasibility of pioglitazone with saroglitazar in patients who have type 2 diabetes mellitus.	At week 24, each therapy group's HbA1c level decreased statistically significantly from baseline, with a p-value of 0.016.	[4]
SPSS (ANOVA)	To ascertain the relationship between lipid profile and IS risk in Patients with t2dm and apolipoprotein E gene polymorphism	Lipedema was associated with a low frequency of diabetes (2%), dyslipidemia (11.7%), and hypertension (13%)	[5]
SPSS (*t*-test)	The test and HBA1c analyses were performed following accepted clinical laboratories, and the data collected represented the outcomes of the most recent blood biochemical analysis performed within the previous three months.	Predictive indicators that could be utilized to indicate treatment in T2DM patients can be seen in lipid profiles (LDL-C) and lipid ratios (LDL-C/HDL-C and TC/HDL-C ratio).	[6]
Machine learning	They assess 35 different machine learning methods and present a classification of diabetes risk factors.	With and without feature selection, the Bagging-LR algorithm provides the most efficiency for a balanced dataset.	[7]
Machine learning	They proposed four machine learning techniques to address the issue of diabetic safety.	The resulting approach drastically lowers the frequency of hypoglycemia events, increasing safety and giving patients more assurance in their decision-making.	[8]

process the data, as seen in Table 1 above. There is non-technical research on data analysis utilizing SEM-PLC. This indicates a lack of sufficient understanding of the related and reliant impacts of the application of risk variables. This supports the necessity for additional research to be added to the empirical studies that looked at the antecedent elements.

3. Methodology

This study used a systematic literature review which is a qualitative method similar to a document review that is appropriate for identifying, evaluating, and interpreting available information about a research topic [11]. A systematic review of the literature (SLR) is conducted to investigate the elements of type 2 diabetes risk factors. Search phrases at this point included "Type 2 diabetes," "Type 2 diabetes and AI," and "Type 2 diabetes factors," among others. IEEE Xplore, Science Direct, Springer, Web of Science, Pub-Med, and Google Scholar were the resources studied. The studies were between 2018 and 2021, these works are either published as journal articles or in conference proceedings. Only 24 of the 47 publications were found to be matched the inclusion standards for this study and only 12 of those addressed the study's research aim. Due to the exclusion of duplicate and irrelevant research based on the theme and the search terms, a structured questionnaire will be employed in conjunction with both qualitative and quantitative methodologies.

Several types 2 diabetes-related variables have been incorporated into the conceptual model to also be evaluated for prediction and to determine whether or not the variables shape patients' recovery according to their blood tests, and which factor does have the biggest effect on developing type 2 diabetes that can be identified. The population of this study is patients in the city of Basrah with type 2 diabetes or who are likely to develop type 2 diabetes. The sample size determined by Roscoe's theory is approximately 10 times the variables or indicators included [9]. Most behavioral studies, according to Roscoe, require a sample size greater than 30 but less than 500. Since this study uses four variables which are Body Mass Index (BMI), triglyceride (TG), hemoglobin (A1C), and total cholesterol (TC), therefore, 200 was set as the sample number. The blood test and survey were collected from various Medical labs in Basrah city. A questionnaire is used in the quantitative approach to reach more respondents who can contribute to this study the questionnaire was adapted from various sources, and the collected data has been analyzed using SEM-PLC software. Fig. 1 shows the steps used in the methodology.

4. Results and discussions

4.1. Findings of the systematic literature review

As observed, the preliminary study showed that lifestyle, medical condition, hereditary, psychosocial, and demographic are core risk factors to be considered in Type 2 diabetes.

4.1.1. Lifestyle factor

The key stakeholders' viewpoints on Type 2 diabetes stated that (a) Smoking, (b) Lack of physical activity, (c) Alcohol, and (d) Environment are the emerging factors. Cigarettes and a lack of exercise were recognized as the main moderators of the rising insulin levels in people who have a low socioeconomic background in an Australian study of health behaviors [10,11]. People with smoking habits can show the ability to have T2D when compared with people who do not smoke [12]. Fig. 2 shows the Lifestyle factor and elements identified by the systematic literature review (SLR).

4.1.2. Medical condition factor

The perspectives of the critical stakeholders for Type 2 diabetes revealed that (a) Obesity, (b) Cardiovascular disease, (c), and High blood pressure are emerging factors [7,13]. Fig. 3 shows the Medical Condition factor and elements identified by the systematic literature review (SLR).

4.1.3. Hereditary factor

The perspectives of the critical stakeholders for Type 2 diabetes revealed that (a) Family History and (b) Ethnicity are the emerging factors. Genetic factors interact to determine impeded-cell insulin secretion and peripheral insulin resistance [14]. Race influences the susceptibility of people with colored skin to develop diabetes [15,16], and this is due to ethnic differences in food consumption, as certain races consume more of certain food groups than others. Furthermore, rates can vary depending on ethnicity. Fig. 4 shows the Hereditary factor and

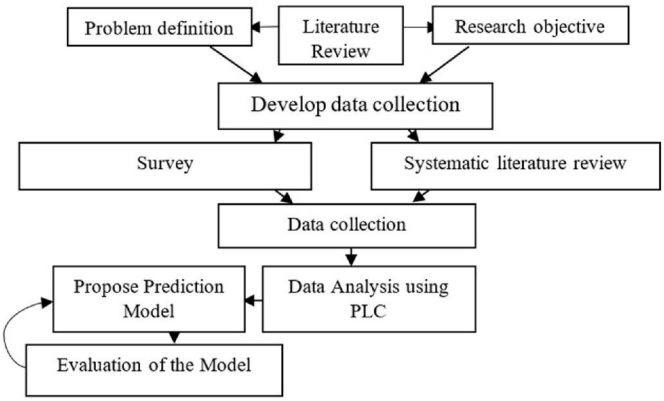

Fig. 1. The methodology steps.

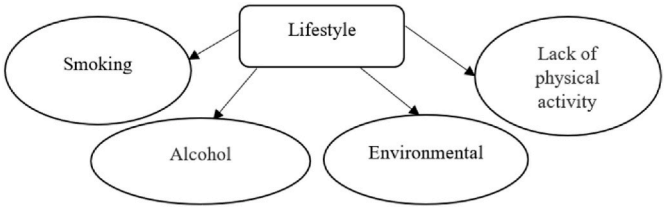

Fig. 2. Lifestyle risk factors and elements identified by the systematic literature review.

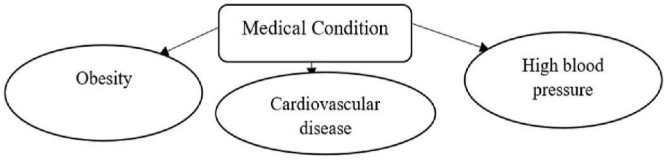

Fig. 3. Medical condition risk factors and elements identified by the systematic literature review.

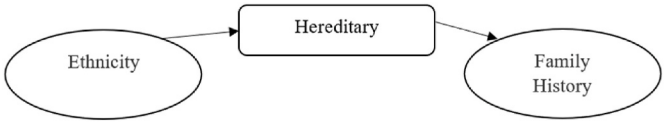

Fig. 4. Hereditary risk factors and elements identified by the systematic literature review.

elements identified by the systematic literature review (SLR).

4.1.4. Psychosocial factor

The perspectives of the critical stakeholders for Type 2 diabetes revealed Stress as a factor, and illness related to an individual's mental health is one example of a psychosocial factor. Patients' ability to cope with chronic illnesses, which are common as age-related disorders, may be influenced by psychological factors. Anxiety, and depression, may impair compliance and adherence, resulting in predictable outcomes and predicting morbidity and mortality independent of other confounders [17]. Fig. 5 shows the Psychosocial factor and elements identified by the systematic literature review (SLR).

4.1.5. Demographic factor

The perspectives of the critical stakeholders for Type 2 diabetes revealed that (a) Age and (b) Gender are the emerging factors. Individual characteristics are referred to as demographic risk factors. In terms of age, there is evidence that as people get older, the overall percentage of type 2 diabetes patients is increasing [20, 21]. However, there is a clear drop-off at some point where men are more likely than women to develop Type 2 Diabetes. Fig. 6 shows the Demographic factor and elements identified by the systematic literature review.

4.2. Findings from the survey

The questionnaires were distributed as many as possible but only 200 responses were used in this research. The results showed that 52% of the

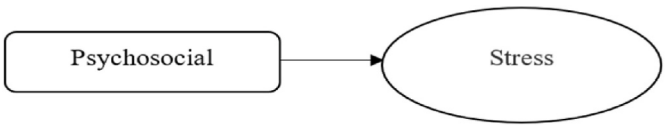

Fig. 5. The psychosocial risk factor and element identified by the systematic literature review.

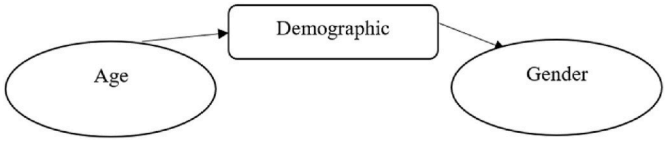

Fig. 6. The demographic risk factor and elements identified by the SLR.

respondents were Male, whereas 38% of the respondents were between the ages of 40 and 49. The majority of the respondents (90%) are married. Table 2 below shows the results received to identify factors and elements related to type 2 diabetes.

According to the survey results, it shows that the majority of patients have depression, which been asked in the questionnaire in terms of feeling sad, anxious, or empty, feelings of despair, feelings of guilt or worthlessness, difficulty falling asleep or staying asleep for long periods, loss of appetite or overeating, loss of interest in activities and hobbies, in addition, those patients were married this also indicates that marital status can be an important element of developing type 2 diabetes. Moreover, the results indicate that the majority of patients have diabetic blood relatives and high blood pressure. In addition, they do not perform any physical activities.

4.3. Factor analysis of the dataset

The factor analysis technique is used for breaking down large variables into a smaller number of factors. From each variable, the most common variations are taken out and added to the common variables. The analysis that follows makes use of these typical variables. The correlation among variables in the form of the matrix is presented in Table 3.

The findings revealed that the correlation is stabilized. The values between single and itself are always one. In the correlation, the principal diagonal is 1.000 so each variable has the perfect positive linear relationship with itself. The calculation is shown in Table 3 in which the overall root square value of AVE surpasses the inter-latent variable correlation rating, and the P-value is 0.05. The P-values correlations for each variable are presented in Table 4.

The loading factor, composite reliability, average variance, and discriminant validity analysis are the foundations for the outer model measurement of variables with reflective indicators. The findings of the loading factor analysis are shown in Table 5.

Because the P-value is 0.05, the relationship between latent variables with their indicators is strong. The variable coefficients are shown in Table 6.

The R-squared showed that there is a strong relationship between Body Mass Index (BMI with hemoglobin (A1C) and triglyceride (TG). In addition, according to Henseler et al. [18], In the PLS-SEM analysis, the lowest composite reliability value should be greater than 0.7. The measurements of scale items showed good composite reliability, which

Table 2
Survey results.

Elements	Survey results (respondent)
Age (from 40 to 49)	76
Gender	105 are Male
Level of education	124 University graduated
Marital Status	180 married
Place of residence	120 City centers
Eating vegetables or fruit	101 dailies
Physical activity	137 do not perform any physical activity
Smoke	155 do not smoke
High blood pressure	108 with high blood pressure
Depression	180 with depression
Blood relatives with diabetes	163 has type 2 diabetes with a relative in the family

Table 3
Correlations among l.vs. with sq. rts. of AVEs.

	BMI	A1C	TC	TG
BMI	(1.000)	0.280	0.349	0.254
A1C	0.280	(1.000)	0.128	0.223
TC	0.349	0.128	(1.000)	0.027
TG	0.254	0.223	0.027	(1.000)

Table 4
P- values for correlations.

	BMI	A1C	TC	TG
BMI	1.000	0.003	<0.001	0.008
A1C	0.003	1.000	0.183	0.019
TC	<0.001	0.183	1.000	0.782
TG	0.008	0.019	0.782	1.000

Table 5
The indicator loading and cross-loading.

	BMI	A1C	TC	TG	P-Value
BMI	(1.000)	0.280	0.349	0.254	<0.001
A1C	0.280	(1.000)	0.128	0.223	<0.001
TC	0.349	0.128	(1.000)	0.027	<0.001
TG	0.254	0.223	0.027	(1.000)	<0.001

Table 6
Variable coefficients.

	BMI	A1C	TC	TG
R-squared		0.079	0.122	0.064
Composite reliable	1.000	1.000	1.000	1.000
Cronbach's alpha	1.000	1.000	1.000	1.000
Full Collin. VIF	1.277	1.117	1.146	1.104
Min	−3.311	−0.995	−1.191	−3.530
Max	2.004	4.925	2.515	2.380
Median	0.068	−0.360	0.075	−0.155
Mode	−3.311	−0.584	−1.191	−0.530
Mode	−0.983	2.414	0.152	0.076

is (1.000) and Cronbach's alpha is also 1.000 which indicates a very good level of reliability. Moreover, Full Collin. VIF Collinearity statistics are (1.000) which measures the amount of multicollinearity among a set of multiple regression variables and it should be less than 4 to be accepted. Bartlett's test of Sphericity was conducted which is a test where the important value of that test is one being the significance level that is smaller than 0.001. The null hypothesis of Bartlett's test can be rejected because its smaller than 0.000. Therefore, the requirement is met due to the significant level being smaller than 0.01. Table 7 shows the KMO and Bartlett's test.

To determine statistical significance, this study utilizes a measurement known as the p-value: if the p-value is lower than the significance level, the result is statistically significant. The indicator loading and cross-loading along with the KMO and Bartlett's test have proven that the analysis of the data is significant.

This study designed a prediction Model by using a PLC-regression outer model analysis algorithm. The percentage of R2 variance of the endogenous latent variable on the exogenous variables shows that Body Mass Index (BMI) strongly influences hemoglobin (A1C) with R2 of 80%, whereas Body Mass Index (BMI) strongly influences triglyceride (TG) with R2 of 60%. Triglycerides (TG) and hemoglobin (A1C) were the most significant risk variables for type 2 diabetes, as illustrated in Fig. 7.

Table 7
The KMO and Bartlett's test.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.582
Bartlett's Test of Sphericity	Approx. Chi-Square	135.378
	Sig.	.000

The significant level in SEM- PLS analysis should be less than 0.05. Therefore, when the obtained p-value or R^2 value are clustered near the central location (near 0.0) the value will be more significant whether it is positive or negative [19]. In our prediction model, there are three hypotheses one for each variable. Therefore, the prediction model should have three p-values as shown in Fig. 7 and all variables' p-value should be < .01. According to Yong et al. [20], who recommended that R^2 values should be (equal to or greater) than 0.10 for the variance of a particular endogenous construct to be deemed adequate. In our model, this value is represented in the first hypothesis, which is the relationship between BMI and TC in this relationship $R^2 = 0.12$ which is better than the second hypothesis where $R^2 = 0.08$ and the third hypothesis where $R^2 = 0.06$. Fig. 7 shows that hemoglobin (A1C) and triglyceride (TG) have met the requirement, and are acceptable.

5. Conclusion

The risk factors and elements associated with type 2 diabetes have been identified in this study. Factors risk which develops type 2 diabetes has been examined due to their importance in predicting diabetes for early diagnosis and therapy. Triglyceride (TG) and hemoglobin (A1C) have been identified in this study as the most influencing factors for developing type 2 diabetes. Different analysis techniques have been used to determine the significance of data analysis where the significant level in SEM-PLS is less than 0.05. However, in this study, factors that influence the development of type2 diabetes are limited and may vary from one patient to another. The other limitation of this study is the sample size compared to the patient population in Iraq.

CRediT authorship contribution statement

Waleed Noori Hussein: Conceptualization, Methodology, Study design, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Zainab Muzahim Mohammed:** Conceptualization, Methodology, Study design, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Amani Naama Mohammed:** Conceptualization, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial

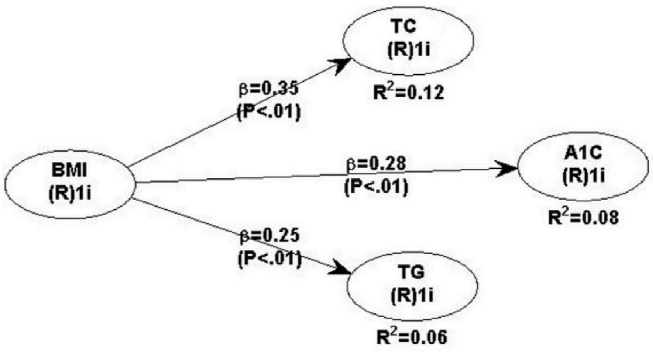

Fig. 7. Prediction model.

interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] M.D. Campbell, et al., Benefit of lifestyle-based T2DM prevention is influenced by prediabetes phenotype, Nat. Rev. Endocrinol. 16 (7) (2020) 395–400, https://doi.org/10.1038/s41574-019-0316-.

[2] D.M. Tanase, et al., The intricate relationship between type 2 diabetes mellitus (T2DM), insulin resistance (IR), and nonalcoholic fatty liver disease (NAFLD), 1, J. Diabetes Res. 20 (20) (2020), 202016, https://doi.org/10.1155/2020/3920196.

[3] R. Chawla, S. Madhu, B. Makkar, S. Ghosh, B. Saboo, S. Kalra, RSSDI-ESI clinical practice recommendations for the management of type 2 diabetes mellitus 2020, Indian J. Endocrinol. Metabol. 24 (1) (2020) 1–5, https://doi.org/10.4103/ijem.IJEM_225_20.

[4] M. Krishnappa, et al., Effect of saroglitazar 2 mg and 4 mg on glycemic control, lipid profile and cardiovascular disease risk in patients with type 2 diabetes mellitus: a 56-week, randomized, double blind, phase 3 study (PRESS XII study), Cardiovasc. Diabetol. 19 (1) (2020) 1–13, https://doi.org/10.1186/s12933-020-01073-w.

[5] N.P. Maratni, et al., Association of apolipoprotein E gene polymorphism with lipid profile and ischemic stroke risk in type 2 diabetes mellitus patients, J. Nutr. Metabol. 20 (21) (2021) 1–16, https://doi.org/10.1155/2021/5527736.

[6] B.A. Artha Imjr, N.K. Dharmawan, U.W. Pande, K.A. Triyana, P.A. Mahariski, J. Yuwono, V. Bhargah, I.P.Y. Prabawa, I.B.A.P. Manuaba, I.K. Rina, High level of individual lipid profile and lipid ratio as a predictive marker of poor glycemic control in type-2 diabetes mellitus, Vasc. Health Risk Manag. 5 (2019) 149–157, https://doi.org/10.2147/VHRM.S209830, doi: 10.2147/VHRM.S209830.

[7] L. Ismail, H. Materwala, M. Tayefi, P. Ngo, A.P. Karduck, Type 2 diabetes with artificial intelligence machine learning: methods and evaluation, Arch. Comput. Methods Eng. 29 (1) (2022) 313–333, https://doi.org/10.1007/s11831-021-09582-x.

[8] J. Vehí, I. Contreras, S. Oviedo, L. Biagi, A. Bertachi, Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning, Health Inf. J. 26 (1) (2020) 703–718, https://doi.org/10.1177/1460458219850.

[9] U. Sekaran, R. Bougie, Research Methods for Business: A Skill Building Approach, john wiley & sons, 2016.

[10] S.J. Carroll, M.J. Dale, T. Niyonsenga, A.W. Taylor, M. Daniel, Associations between area socioeconomic status, individual mental health, physical activity, diet and change in cardiometabolic risk amongst a cohort of Australian adults: a longitudinal path analysis, PLoS One 15 (5) (2020), https://doi.org/10.1371/journal.pone.0233793.

[11] H. Kolb, S. Martin, Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes, BMC Med. 15 (1) (2017) 1–11, https://doi.org/10.1186/s12916-017-0901-x.

[12] X. Wei, E. Meng, S. Yu, A meta-analysis of passive smoking and risk of developing Type 2 Diabetes Mellitus, Diabetes Res. Clin. Pract. 107 (1) (2015) 9–14, https://doi.org/10.1016/j.diabres.2014.09.019.

[13] C. King, M.A. Lanaspa, T. Jensen, D.R. Tolan, L.G. Sánchez-Lozada, R.J. Johnson, Uric acid as a cause of the metabolic syndrome, Uric Acid. Chron. Kidney Dis. 192 (2018) 88–102, https://doi.org/10.1159/000484283.

[14] Q. Ge, X. Xie, X. Xiao, X. Li, Exosome-like vesicles as new mediators and therapeutic targets for treating insulin resistance and β-cell mass failure in type 2 diabetes mellitus, J. Diabetes Res. 20 (19) (2019), https://doi.org/10.1155/2019/3256060.

[15] N.A. Werissa, P. Piko, S. Fiatal, Z. Kosa, J. Sandor, R. Adany, SNP-based genetic risk score modeling suggests no increased genetic susceptibility of the Roma population to type 2 diabetes mellitus, Genes 10 (11) (2019) 942, https://doi.org/10.3390/genes10110942.

[16] N. Seshagiri Rao, K. Kalyani, B. Mitiku, Fixed point theorems for nonlinear contractive mappings in ordered b-metric space with auxiliary function, BMC Res. Notes 13 (1) (2020), https://doi.org/10.1186/s13104-020-05273-1.

[17] G. Martino, V. Langher, V. Cazzato, C.M. Vicario, Psychological factors as determinants of medical conditions, 10, Front. Media. SA.. 19 (25) (2019) 777–780, https://doi.org/10.3389/fpsyg.2019.02502.

[18] J. Henseler, et al., Common beliefs and reality about PLS: comments on rönkkö and evermann, Organ. Res. Methods 17 (2) (2014) 182–209, https://doi.org/10.1177/1094428114526928.

[19] J.F. Hair, J.J. Risher, M. Sarstedt, C.M. Ringle, When to use and how to report the results of PLS-SEM, Eur. Bus. Rev. 31 (1) (2019) 2–24, https://doi.org/10.1108/EBR-11-2018-0203, 2019.

[20] J.Y. Yong, M. Yusliza, T. Ramayah, O. Fawehinmi, Nexus between green intellectual capital and green human resource management, J. Clean. Prod. 215 (2018) 364–374, https://doi.org/10.1016/j.jclepro.2018.12.306.