BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Factors affecting the survival of prediabetic patients: comparison of Cox proportional hazards model and random survival forest method

Mehdi Sharafi[1], Mohammad Ali Mohsenpour[2,3], Sima Afrashteh[4*], Mohammad Hassan Eftekhari[3], Azizallah Dehghan[8], Akram Farhadi[5], Aboubakr Jafarnezhad[2], Abdoljabbar Zakeri[6] and Mehdi Azizmohammad Looha[7]

Abstract

Background The worldwide prevalence of type 2 diabetes mellitus in adults is experiencing a rapid increase. This study aimed to identify the factors affecting the survival of prediabetic patients using a comparison of the Cox proportional hazards model (CPH) and the Random survival forest (RSF).

Method This prospective cohort study was performed on 746 prediabetics in southwest Iran. The demographic, lifestyle, and clinical data of the participants were recorded. The CPH and RSF models were used to determine the patients' survival. Furthermore, the concordance index (C-index) and time-dependent receiver operating characteristic (ROC) curve were employed to compare the performance of the Cox proportional hazards (CPH) model and the random survival forest (RSF) model.

Results The 5-year cumulative T2DM incidence was 12.73%. Based on the results of the CPH model, NAFLD (HR = 1.74, 95% CI: 1.06, 2.85), FBS (HR = 1.008, 95% CI: 1.005, 1.012) and increased abdominal fat (HR = 1.02, 95% CI: 1.01, 1.04) were directly associated with diabetes occurrence in prediabetic patients. The RSF model suggests that factors including FBS, waist circumference, depression, NAFLD, afternoon sleep, and female gender are the most important variables that predict diabetes. The C-index indicated that the RSF model has a higher percentage of agreement than the CPH model, and in the weighted Brier Score index, the RSF model had less error than the Kaplan-Meier and CPH model.

Conclusion Our findings show that the incidence of diabetes was alarmingly high in Iran. The results suggested that several demographic and clinical factors are associated with diabetes occurrence in prediabetic patients. The high-risk population needs special measures for screening and care programs.

Keywords Survival analysis, Prediabetic, Cox regression model, Random Survival Forest, Afternoon sleep, Depression, Non-alcoholic fatty liver disease

*Correspondence:
Sima Afrashteh
sima.afrashte3@gmail.com

Full list of author information is available at the end of the article

Introduction

The prevalence of type 2 diabetes mellitus (T2DM) among adults is rapidly increasing worldwide [1]. After cancers and cardiovascular diseases, T2DM is estimated to be the third chronic disease that has become one of the leading public health challenges [2]. It affects 463 million individuals globally, and this number is predicted to increase to 700 million by 2045 [3]. Of individuals suffering from T2DM, 80% live in low- and middle-income countries [4]. Thus, the average prevalence of T2DM in the Eastern Mediterranean Region (EMRO) among adults was estimated to be 13.7%, which is the highest rate compared to other WHO regions. In Iran, the prevalence of this disease in adults is 11.9% [5].

Prediabetic patients, defined by fasting blood sugar of 100 to 125 mg/dl, are at high risk for T2DM. High blood glucose levels, weight gain, high plasma insulin levels, dyslipidemia, hypertension, and decreased beta-cell function are named risk factors for T2DM in pre-diabetes [6]. Diabetes can cause damage to various tissues, especially the eyes, kidneys, heart, blood vessels, and nerves, through micro- and macro-vascular complications [7]. Research shows that adopting a healthy lifestyle, including proper physical activity, a healthy diet, and medication use (such as metformin), can prevent disease progression in high-risk individuals [8].

Survival analysis is a statistical approach where the outcome variable of interest is the time until an event occurs [9]. The most popular and widespread survival analysis method in medicine is the CPH model, which shows the importance of variables using the Hazard ratio (HR); however, this model limits assumptions such as the proportionality of hazard [9, 10]. Also, this model is invalid in conditions with high censorship [11]. Therefore, using models that predict risk factors with fewer assumptions is necessary. In recent years, new methods have been proposed for this analysis, including machine learning, a complete class of these techniques, and making fewer assumptions about the use of data [12]. Random Survival Forest (RSF) is a non-parametric machine learning method used to solve the problem of using the CPH model [11]. Also, it is known that the CPH model is a semi parametric method in which survival times are assumed about predictor variables in a particular way and proportional hazards are assumed [13]. In contrast, RSF models are useful for the discovery of knowledge because using a partial dependence plot (PDPs) approach to visualize relationships between variables is helpful in exploring complex relationships between variables [14].

Due to the increasing prevalence of T2DM, various models have been used for the early detection of the disease, including the CPH and RSF models [15]. Although the proper performance of RSF compared to the CPH model has been shown in various studies [16, 17], to our knowledge, no study has been conducted to use this model in predicting diabetes in prediabetic individuals. Therefore, in the present study, we compared RSF and CPH models to identify a robust approach to identifying risk factors for diabetes in prediabetic individuals in south of Iran.

Method

Study design

This is a prospective cohort study, a sub-branch of the Fasa Persian cohort study (FACS). The Fasa Persian cohort was started in 2016, and its primary goal is to investigate the factors affecting the incidence of cardiovascular diseases; so far a 5-year follow-up period have been implemented. Individuals within the age range of 35–70 years were planned to participate as the target population in the cohort. Fasa city, with a population of approximately 250,000, is situated in the eastern region of the Fars province in southwest Iran. A rural region known as Sheshdeh (with a total population of 41,000 was chosen for the cohort study [18]. This study was under the Helsinki Declaration and the study protocol was approved by the institutional review board and ethics committee of Fasa University of Medical Sciences (IR.FUMS.REC.1401.007).

In the Fasa cohort study, comprehensive questionnaires to record demographic variables, socioeconomic status, history of communicable and non-communicable diseases, anthropometric measurements, physical examinations, blood pressure and pulse, nutritional status, and blood and urine analysis have been used for the baseline assessments. Also, blood, hair, and nail samples have been collected in a biobank for possible subsequent investigations. Trained interviewers have completed questionnaires through interviews. The FASA cohort study protocol with the study procedure and detailed information has been published previously [18].

Study population

In the present study, the baseline information of the FACS has been extracted to identify prediabetic patients. Based on the definition of the American Diabetes Association, individuals with fasting blood sugar (FBS) between 100 and 125 mg/dl were considered prediabetic [19, 20]. These patients have been examined for the incidence of diabetes during 5 years of annual follow-up. The study interviewers, who are nurses, conducted the follow-up of these patients. The confirmation of diabetes was carried out based on the diagnostic criteria outlined by the American Diabetes Association, which includes: Glycated hemoglobin (A1C≥6.5%.) or Fasting plasma glucose FPG≥126 mg/dL (7 mmol/L). Fasting is defined as no caloric intake for at least 8 h or 2-hour plasma glucose≥200 mg/dL during an oral glucose tolerance test

(OGTT). The test should be performed as described by the World Health Organization, using a glucose load containing the equivalent of 75 g of anhydrous glucose dissolved in water.

Or in a patient with classic symptoms of hyperglycemia or hyperglycemic crisis, random plasma glucose ≥ 200 mg/dL [21].

Variables of the study

The dependent variable for the current study was the incidence of diabetes in people with pre-diabetes and the time to the incidence of diabetes (months) within five years. Also, the variables including age, gender (male/female), education status (illiterate or literate), marital status (married or single/divorced/widowed), waist circumference: was measured by wrapping a tape measure around the body at the level of the navel, halfway between the lowest rib and the top of the hipbone, Waist to Hip Ratio: (WHR) is calculated by dividing the waist measurement by the hip measurement, fasting blood glucose (FBS), afternoon sleep (hour), body mass index (BMI) is calculated by dividing a person's weight in kilograms by the square of their height in meters. (underweight: < 18.5; normal: 18.5–24.9; overweight: 25–29.9; obese > 29.9, kg/m^2), The self-declared past medical history: non-alcoholic fatty liver disease (NAFLD), and suffering from depression (Confirmation by a physician and taking medication) have been investigated as independent variables. All independent variables have been measured for all participants at the recruitment.

Definitions of the methods

In survival analysis, various regression models are used for predicting the probability of incidence of future events. The Cox Proportional Hazards (CPH) model is one of the most widely used survival models that examine the time to event and the factors affecting it. An essential assumption of the CPH model is the Cox proportional hazards assumption, which is important for all independent variables in the CPH model. One of the ways to check this assumption is the Schönfeld residuals. The proportional hazard assumption is supported by a non-significant relationship between residuals and time [22]. It provides Hazard Ratios (HR) and p-values for covariates, making it interpretable in terms of risk factors affecting survival [23]. One of the statistical models used to analyze time-to-event data, especially when there is the right sensor, and events are time-dependent, is Random Survival Forests (RSF). This model uses a set of decision trees to rank and predict important variables that affect the event [16]. It excels in handling complex interactions and nonlinear relationships in the data, making it robust for survival prediction [23, 24]. In order to estimate the importance of each covariate, the forest is first grown with the real data and then with the permuted data on the desired covariate. The difference in the accuracy of predictions in these two conditions determines the importance of the variable. After selecting the best model, to estimate each variable's importance, the variables were first sorted and then the covariates whose value was zero and negative were removed. because these variables have no ability to predict the dependent variable, while the variables with positive values suggest variables with predictive abilities [11].

Statistical analysis

We have investigated the time to the incidence of T2DM in prediabetic patients and the factors affecting it. In this model, hazard ratio and 95% confidence interval (CI) were used to report the relationship between independent variables and the incidence of diabetes. According to Akaike Information Criterion (AIC), stepwise variable selection is used to finalize the CPH model. Stepwise variable selection is an iterative algorithm. This method alternates between forward and backward to achieve a set of stable variables by bringing in and removing variables. This algorithm is more complex and maybe more appropriate than crude (univariate) analysis. Kaplan-Meier plot and 95% confidence interval (CI) were also used to assess the time of incidence of diabetes during 5 years of follow-up. For a more accurate assessment of the factors affecting the survival of prediabetic patients, in addition to the CPH model, we used the Random survival forest (RSF) model with similar variables. In this model, for reporting the associations, the percentage of important variables on the incidence of diabetes was used for reporting. The Concordance index (C-index) and Weighted Brier Score index were used to compare the RSF and CPH model performance. The Concordance Index (C-index) is the most widely used metric in Survival Analysis to evaluate the prediction model. This index can quantify the rank correlation between survival times and risk predictions. It does not depend on selecting a fixed time for assessing the model and specifically takes into account the censoring of persons [25]. Finally, a p-value < 0.05 was considered statistically significant. All the statistical analysis was performed in R software (version 4.2.1). RSF and CPH models were trained using the RandomForestSRC and survival R packages, respectively. Several R packages were used to calculate the integrated log loss, C index, Brier score, and nested cross-validation, including mlr3, mlr3proba, mlr3extralearners, and mlr3pipelines. The pec package illustrated the prediction error curves and concordance index over time. In addition, the "timeROC" package was used for the time-dependent Receiver Operating Characteristic (ROC) curve.

Tune parameters and nested cross-validations

A nested cross-validation approach was used to tune the parameters of the random forest. Accordingly, the mtry and ntrees parameters in this nested cross-validation were set to integer values ranging from 5 to 50 and 64 to 128, respectively, for the random forest, and the ties parameter for Cox was set to be "efron," "breslow," "exact.". Additionally, 10-fold cross-validation was considered for both internal and outer cross-validation. Based on nested cross-validation, three criteria were calculated for random forest and Cox models: C index, integral log loss, and barrier score. The log-loss indicates how close the prediction probability is to the actual/true value (0 or 1 in binary classification). A higher log-loss value means the predicted probability is farther from the actual value.

Variable importance

Two methods were used to develop models for predicting prediabetic probability. First, the train-test method was used to randomly divide the data into two groups, one for training data (70% of samples) and the other for test data (30% of samples), to determine the importance of the variables using RSF. The permutation method was applied with 100 bootstraps resampling to calculate the importance of variables. Further, confidence intervals and standard errors for variable importance were calculated based on the subsampling method using double bootstrapping. Variables with confidence intervals that did not include zero were considered to be selected variables.

In the second step, nested cross-validation was used to compare the performance of both models of CPH and RSF with selected variables. In this method, both inner and outer cross-validation with iteration was used to calculate the C index, Brier score, and integrated log-loss.

Results

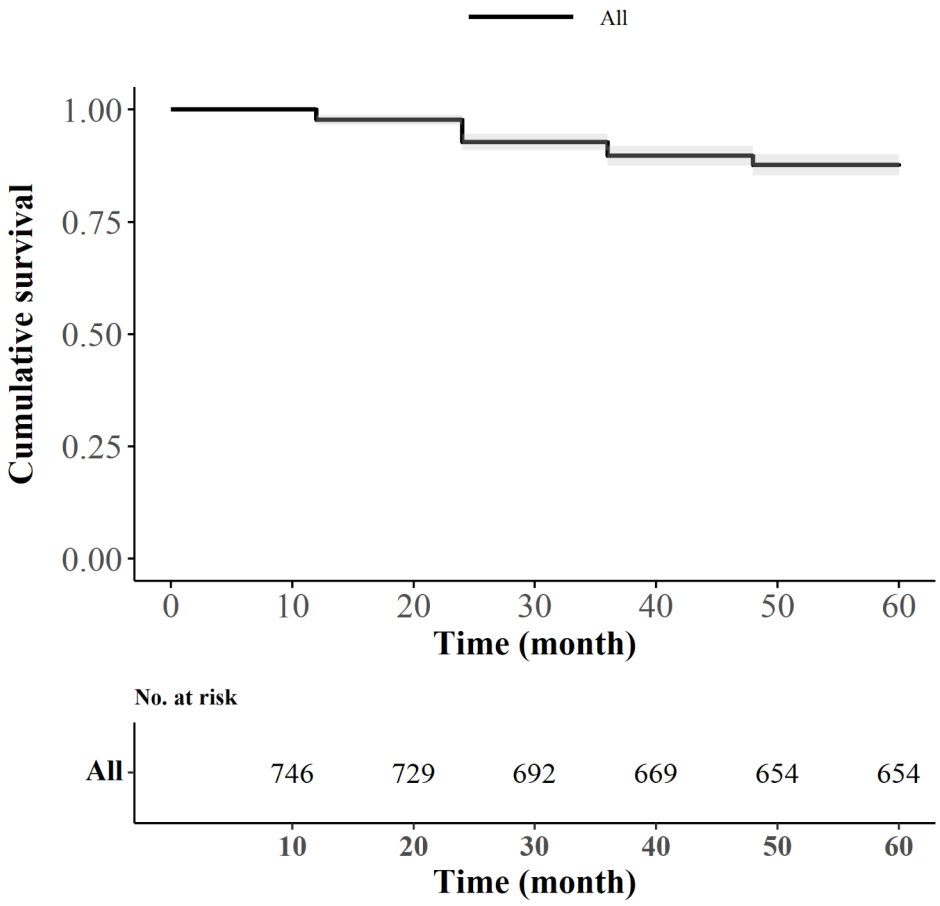
A total of 10,131 individuals participated in the FACS. Of them, 1248 were diagnosed with diabetes at the baseline data, and the prevalence of diabetes was 12.31% (95% confidence interval (CI): 11.68, 12.97). Moreover, the number of people with no diabetes was 8883, of them, 746 had Fasting Blood Sugar (FBS) between 100 and 125 which were classified as prediabetics. The prevalence of prediabetes was 8.39% (95% confidence interval (CI): 7.82, 8.99). The baseline demographic and clinical characteristics of prediabetic patients are shown in Table 1. Based on these results, the mean age of the diabetes patients in the study was reported as 51.98 ± 8.91(mean ± SD) years. Among patients, 56.8% were women, and 50.50% were illiterate. In this study, there was a significant relationship between depression and NAFLD with the incidence of diabetes ($P<0.05$). During 60 months of follow-up in prediabetic patients for incidence of T2DM, 95 pre-diabetes progressed to diabetes. Thus, the 5-year cumulative incidence was 12.73% (95% confidence interval (CI): 10.42, 15.34). The Kaplan-Meier diagram with 95% confidence interval (CI) is depicted in Graph 1.

At first, the unadjusted and adjusted CPH model was fitted. In the CPH model fitting, the Stepwise variable

Table 1 The baseline demographic and clinical characteristics of the study population

Variable	category	Diabetes		*p*-value
		Yes 95(12.7)	No 651(87.3)	
Age(mean ± SD) year		51.98 ± 8.91	52.72 ± 9.26	0.465
Gender n (%)	Male	41(43.2)	266(43.2)	0.738
	Female	54(56.8)	385(59.1)	
Educational status n (%)	Illiterate	48(50.5)	377(57.9)	0.184
	Literate	47(49.5)	274(42.1)	
Marital status n (%)	Married	83(87.4)	552(84.8)	0.643
	Single/divorce/widow	12(12.6)	99(15.2)	
BMI n (%)	Underweight	1(1.1)	27(4.2)	0.028
	Normal	21(22.1)	209(32.2)	
	Overweight	39(41.1)	256(39.4)	
	Obese	34(35.8)	158(24.3)	
Nonalcoholic Fatty Liver n (%)	Yes	23(24.6)	76(11.7)	0.002
	No	72(75.8)	575(88.3)	
Depression n (%)	Yes	9(9.5)	34(5.2)	0.101
	No	86(90.5)	617(94.8)	
Waist Circumference(mean ± SD) cm		102.07 ± 12.43	96.23 ± 12.26	< 0.001
Wrist Circumference(mean ± SD) cm		17.31 ± 1.52	16.88 ± 1.37	0.004
FBS (mean ± SD) mg/dl		129.74 ± 45.47	113.91 ± 25.80	< 0.001
Afternoon Sleep (mean ± SD) hour		1.03 ± 0.97	0.86 ± 0.89	0.081

BMI: body mass index



Graph 1 Kaplan-Meier survival for prediabetic patients during 60 months follow-up in Fasa Persian cohort study

selection method based on Akaike information criteria (AIC) has been used. According to the adjusted results, the hazard of progression of pre-diabetes to diabetes was 74% higher in patients with NAFLD compared to those without NAFLD (HR=1.74, 95% confidence interval (CI): 1.06, 2.85, P=0.028). In addition, increased abdominal fat was associated with a higher risk of diabetes in these groups (HR=1.02, 95% confidence interval (CI): 1.01, 1.04, P=0.001). Also, the risk of progression of diabetes among high fasting blood sugar patients was higher higher compared to people with normal blood sugar (HR=1.008, 95% confidence interval (CI): 1.005, 1.012, P<0.001). Table 2 shows the CPH for modeling the survival of prediabetic patients during 60-month follow-up. The random Survival Forest (RSF) model was used to compare and determine the appropriate model to investigate the factors affecting the incidence of diabetes in pre-diabetes patients. For reporting the results of the RSF model, the importance of variables that were associated with the progression of pre-diabetes to diabetes was used. The results of this model are demonstrated in Graph 2. This analysis found that FBS, waist circumference, depression, NAFLD, afternoon sleep, and female gender are the most important variables for predicting diabetes. These variables were used in the final analysis to develop the minimal and adequate model for predicting pre-diabetes.

Also, the Concordance Index (C-index), Weighted Brier Score, and time -dependent ROC curve were used to compare CPH and RSF models. Results are shown in Graphs 3 and 4, respectively. Based on these indices, as seen in C-index, the RSF model has a higher percentage of agreement than the CPH model, and in the Weighted Brier Score index, the RSF model had less error than the Kaplan-Meier and CPH model. In addition, the ROC curve (AUC: 0.836 vs. AUC:0.683) results also showed a higher area under the curve for the RSF model (Graph 5 and 6).

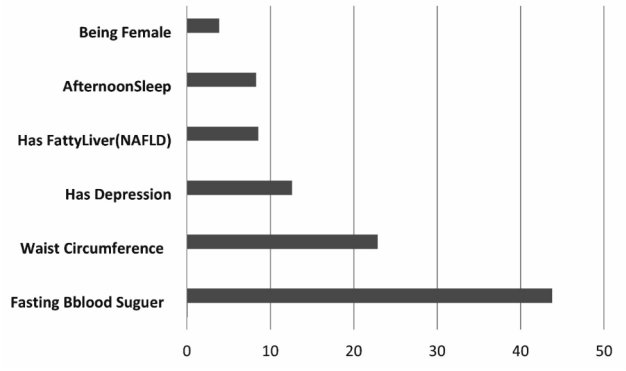
According to Table 3, the integrated log loss of Cox regression and the random forest was 0.222 and 0.113, respectively. In addition, the Brier score for random was 0.046, while it was 0.070 for Cox.

Table 2 CPH model for modeling the survival of prediabetic patients due to 60-month follow-up in the Fasa Persian cohort study (final model)

Variable	Unadjusted Cox regression		Adjusted Cox regression		PH assumption *p*-values	
	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	Single	Global
age	0.99(0.97, 1.01)	0.495	Not included	-		0.235
Sex						
Male	Reference	-	Not included	-		
female	0.92(0.61, 1.38)	0.702				
Wrist Circumference	1.21(1.05, 1.38)	0.005	1.06(0 0.89, 1.25)	0.493	0.423	
Waist Circumference	1.03(1.02, 1.05)	< 0.001	1.02 (1.01, 1.04)	0.001	0.398	
NAFLD positive	2.18 (1.36, 3.49)	0.001	1.74 (1.06, 2.85)	0.028	0.799	
FBS	1.009(1.005, 0.012)	< 0.001	1.008(1.005, 1.012)	< 0.001	0.056	
BMI	1.06(1.03, 1.10)	< 0.001	0.98(0 0.90, 1.07)	0.700		
Depression						
No	Reference	-	Reference	-	0.375	
yes	1.78(0.89, 3.53)	0.100	1.950(0.97, 3.90)	0.059		
AIC	1213.95					
Overall Concordance index	0.715 (SE = 0.025)					

NAFLD: Nonalcoholic fatty liver diseases; FBS: Fasting blood sugar; SE: Standard error; HR: Hazard ratio; CI: confidence interval; PH: proportional hazard; AIC: Akaike information criteria,

P-values < 0.05 is considered significant



Graph 2 Importance variable for progression prediabetic to diabetic patients in 60 month follow-up, Fasa Persian cohort by random survival forest (RSF)

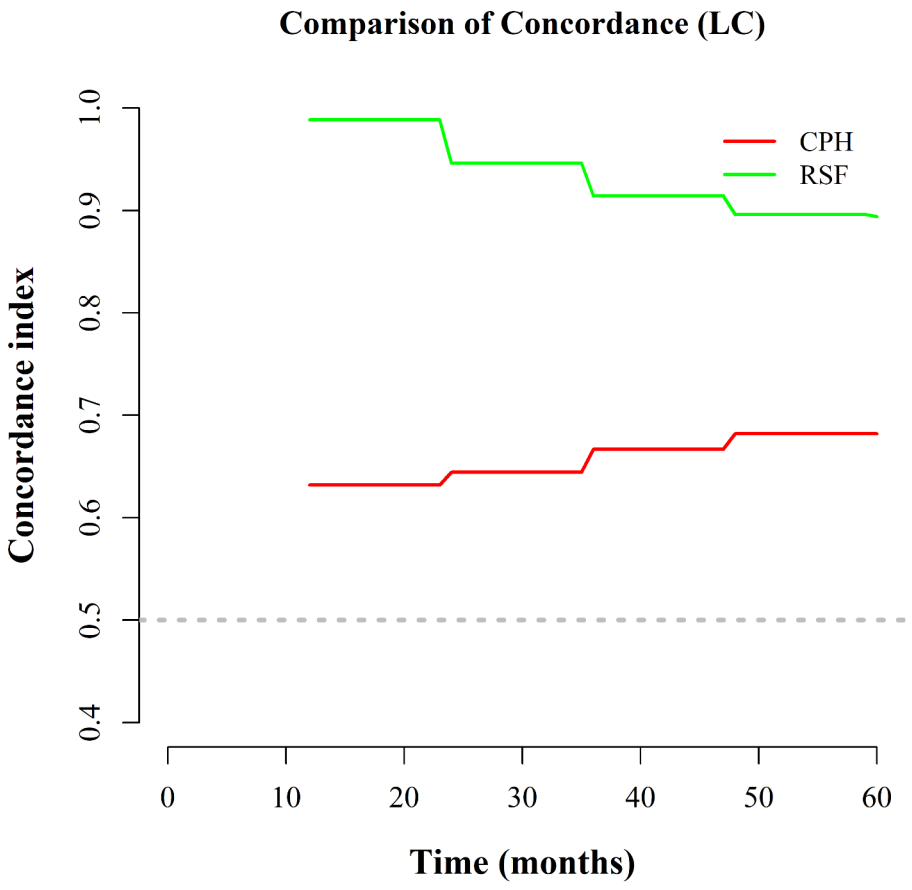
Discussion

This study aimed to compare the RSF model and CPH model to identify risk factors for diabetes in prediabetic individuals in Iran. According to the results, the 5-year cumulative T2DM incidence was 12.7% in prediabetic patients. We have shown that based on the CPH model NAFLD, fasting blood sugar, and increased abdominal fat were directly associated with diabetes occurrence in prediabetic patients. The RSF model suggests that factors including FBS, waist circumference, depression, NAFLD, afternoon sleep, and female gender are the most important variables for predicting diabetes. Also, Based on C-index, the RSF model has a higher percentage of agreement than the CPH model, and in the Weighted Brier Score index, the RSF model had less error than the Kaplan-Meier and CPH model.

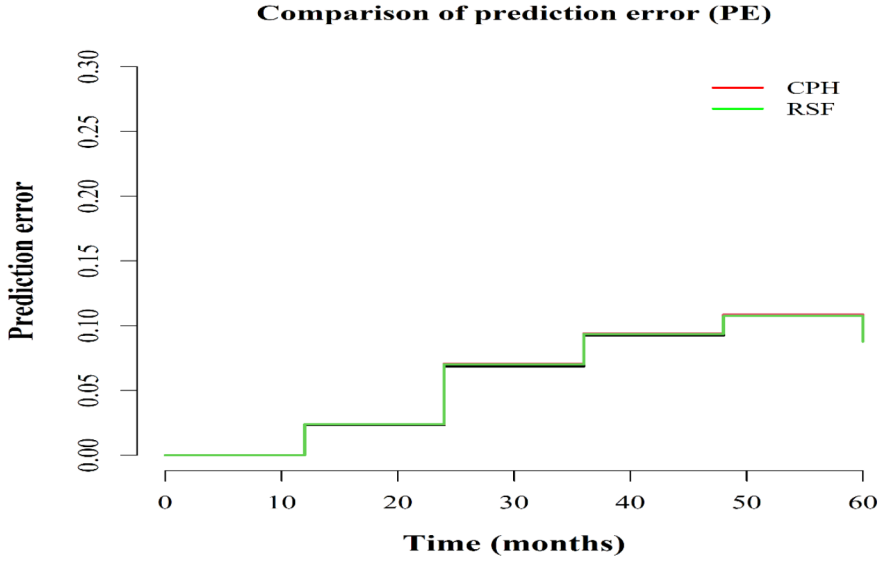
The results of our study showed that the cumulative T2DM incidence in prediabetic individuals was 12.7%. In this longitudinal study of Singapore Malays, age-standardized 6-year cumulative incidence was 11.2% for T2DM, and 20.4% (95% for pre-diabetes [26]. Perhaps the reason for this discrepancy is the longer follow-up duration in Singapore they considered. The present study considers a cumulative incidence for 5 years, while in the Singapore study, it was a period of 6 years. Also, demographic characteristics, genetic factors, lifestyle, medical care, and health interventions can be other influential factors. Studies in Ghana and Spain have estimated a cumulative incidence of between 10.1% and 15.4%, which was also consistent with the results of our study [27, 28].

In the present study, the female gender was an influential factor in increasing the risk of diabetes in prediabetic individuals. The results of studies by Willer et al. [29] confirmed our research findings. Perhaps, it is because women live longer or a type of diabetes called gestational diabetes that makes women more susceptible to diabetes than men [30]. The study conducted by Arnetz et al. [31] also found that the sex factor was effective in diabetes, as found in our study. This may be due to hormonal differences between genders, physiology, and genetic characteristics in men and women [32].

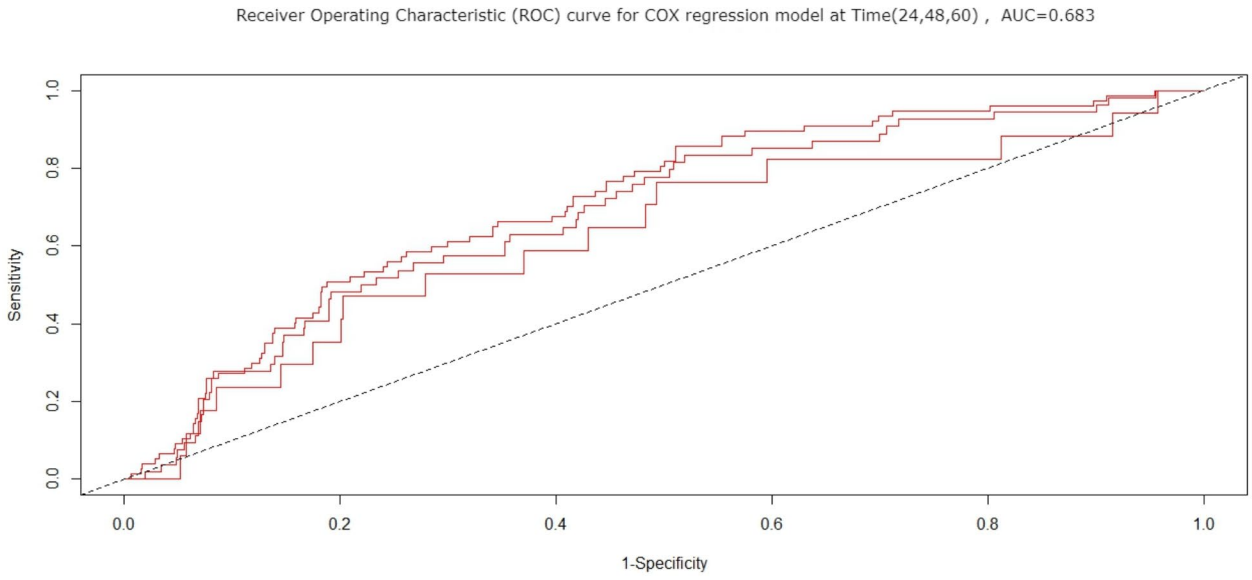
The other risk factor shown to play a role in the progression to type 2 diabetes mellitus is waist circumference. A direct association has been shown between fat accumulations in the body, especially abdominal fat,



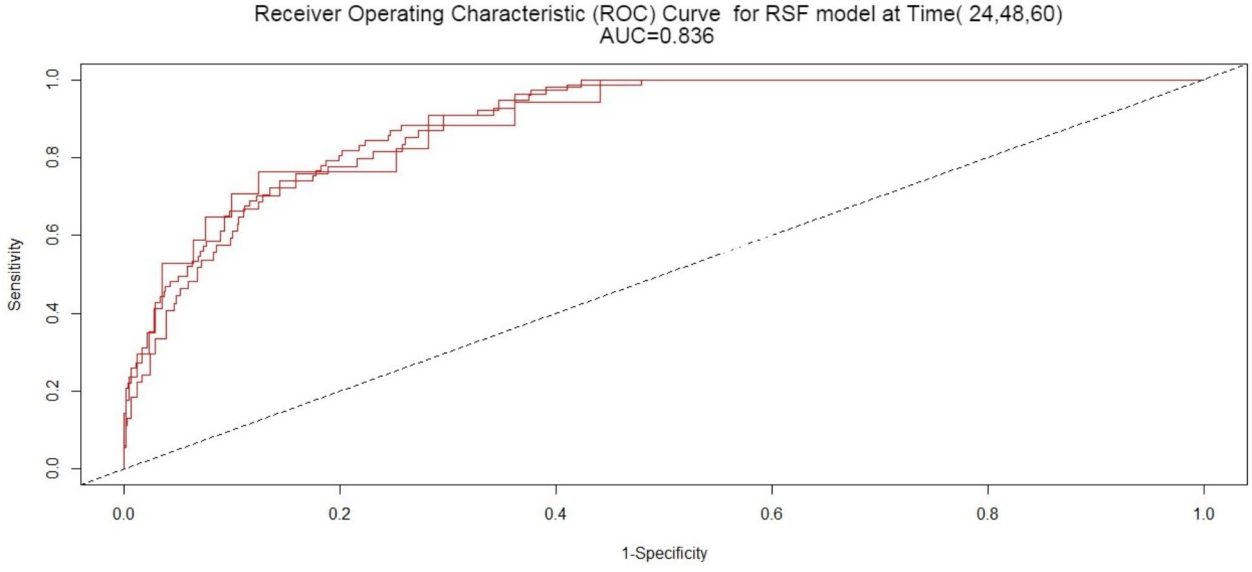
Graph 3 Concordance index (C-index) for comparison of CPH and RSF



Graph 4 Weighted Brier Score for comparison of CPH and RSF



Graph 5 ROC-dependent time for the COX model "AUC: 0.683"



Graph 6 ROC-dependent time for the RSF model "AUC: 0.836"

Table 3 The prediction performance of Cox and random forests

Model	Measures			
	Integrated log loss	C index	Brier score	Area Under the Curve(AUC)
Cox	0.222	0.999	0.070	0.683
Random forests	0.113	0.999	0.046	0.836

and increased waist circumference with diabetes [33, 34]. Increased waist circumferences reflect exceeded fat mass, which causes the production of cytokines, free fatty acids, and inflammatory mediators, which can increase the risk of diabetes [35, 36]. In an investigation of the Korean population, living with increased waist circumference for over 6 years is associated with a higher risk of T2DM than individuals with stable waist circumference [33]. It should be noted that although excess fat mass is related to insulin resistance, accumulation in the abdomen, which is known as high waist circumference, has a stronger association [37]. In the nonalcoholic fatty

liver disease patients, the highest tertile of waist circumference had a 2.04 more chance of showing diabetes [38].

Mental disorders such as depression are found to be a factor that can lead individuals to diabetes. A recent meta-analysis by Graham et al. [39] showed that depression can increase the risk of type 2 diabetes mellitus (RR=1.18). Depressed individuals are prone to change lifestyle factors related to diabetes, including reduced self-care, medication adherence, physical activity, and an increase in following a high-caloric diet, which is deleterious for diabetes and associated with poor glycemic control [40]. Thus, this can lead prediabetic patients toward diabetes.

Of chronic disorders, NAFLD induces diabetes. In fatty liver, majorly, insulin resistance causes type 2 diabetes mellitus. In a previous study, the incidence density of diabetes in patients with non-alcoholic fatty liver disease was 2.27 per 100 people-year, while it was 1.38 in non-NAFLD individuals [38]. In fact, suffering from a fatty liver with different levels of liver fibrosis increases hepatic insulin resistance and causes the release of several pro-inflammatory mediators and pro-diabetogenic hepatokines, which may cause the development of diabetes [41, 42]. However, whether improving or resolving nonalcoholic fatty liver disease can reduce the risk of developing diabetes is controversial. The results of some studies show that the risk of developing diabetes seems to decrease over time after controlling or treating nonalcoholic fatty liver disease [20, 43].

Daytime sleeping was seen to be able to progress the pre-diabetes status toward type 2 diabetes mellitus. A meta-analysis by Guo et al. [44] concluded that longer daytime napping is associated with both a higher prevalence and incidence of diabetes. This increased risk was just seen for sleeping over one hour daily [44]. On the other hand, in comparison with "short night sleep with daytime napping", both "long night sleep with or without daytime napping" had a higher prevalence of diabetes [45]. A disrupted circadian rhythm can be a possible underlying mechanism [46]. By the way, it was shown day time sleep could cause sleep apnea [47], which can trigger a chain of disturbances from oxygen desaturation to increments of catecholamine and cortisol levels. Consequently, this can lead to glucose intolerance [48, 49]. In addition, daytime sleep increases nighttime awakening and shorter nighttime sleep, leading to increased insulin resistance.

Our results indicated higher FBS could increase the risk of progression from pre-diabetes to diabetes. It was shown that various lifestyle modifications and medications could be effective for FBS control. Metformin was observed to reduce the risk of progression to T2DM [50]. In addition, a reduction in calorie intake was proposed to prevent the progression of pre-diabetes to T2DM [50]. Calorie deficit is suggested for waist circumference reduction [51], and NAFLD management [52]. Thus, individuals with pre-diabetes can prevent progression to T2DM with calorie restriction and medication use. In addition, diabetes patients are prone to develop depressive symptoms that affect their diabetes control. Hence it is suggested to follow cognitive behavioral therapies which could lead to better diabetes control [53]. Our finding shows depression can affect diabetes development in Pre-diabetes patients. So, in addition to medical nutrition therapy, lifestyle modification, and medication use, pre-diabetes patients could benefit from psychological therapy to reduce the risk of diabetes.

The results showed that the RSF model has a higher agreement percentage and lower error than the CPH and Kaplan-Meier models in C-index and Weighted Brier Score. In this regard, a study conducted by Safari Et al. Entitled " Identification of Factors Affecting Metastatic Gastric Cancer Patients' Survival Using the Random Survival Forest and Comparison with CPH model also showed that the RSF method, considering it has the highest coordination index and the lowest score, therefore has the highest accuracy and the least error in predicting survival and identification. The most important factors affecting it. This method also has a better performance than the CPH model [54]. A study by Cetin et al. also concluded that the RSF model works better than the CPH model. Perhaps because Such as the unlikeliness of data being associated with a complete set of independent variables, the ease of judging the importance of variables in order to select a variable, as well as its ability to combine nonlinear and interactive roles of multiple variables [12]. Which confirms our research findings. According to Morsy et al., While the CPH model cannot automatically detect the nonlinear effects of all variables and has more errors, RSF models can and do have fewer errors than CPH models [55], which is in line with the results of our study. A study by Karimi et al. showed that according to the C-index, the RSF model has a higher agreement than the CPH model and the Weighted Brier Score; the RSF model had less error than the CPH models [11].

these models were chosen for their distinct characteristics. The CPH model was selected for its interpretability, providing insights into the significance of various prognostic factors [23, 56]. In contrast, the RSF model was chosen for its ability to handle complex data patterns and interactions without assuming proportional hazards. This makes it suitable for capturing intricate relationships between variables that may not follow a linear or proportional trend [23, 57].

Strengths and limitations: Among the strengths of this study are the sufficient sample size and study population to determine the role of independent variables in the incidence of diabetes, we used the Random Forest model

in addition to the traditional CPH model. One of the limitations of the present study is that we only used the information of people aged 35 to 70 years.

Conclusion

The present study evaluated the factors affecting the survival of Prediabetic Patients and their related factors. Our results suggested that several demographic and clinical factors, including NAFLD, FBS, high abdominal fat, waist circumference, depression, evening sleep, and female gender, are significantly associated with diabetes in prediabetic patients. These findings emphasize the need to understand the factors associated with pre-diabetes in Iran so that early effective and lifestyle interventions can be implemented. In addition, Screening, and health promotion activities, including dissemination of information in the clinics, mass media, community events, and proper management of pre-diabetes might contribute to primary and secondary prevention of diabetes.

Acknowledgements
We appreciate all the participants who participated in the study.

Author contributions
MSh: final analysis, providing the main idea of the study, and methodology; MAM, MHE, AF and AD: writing the manuscript, developing the idea, and revising the final manuscript; SA: final analysis, supervised the study and revising the final manuscript; AJ and AZ: contributing to data analysis; MAL: revising the final manuscript. All authors approved the final version of the manuscript.

Funding
Fasa University of Medical Sciences, Fasa, Iran financially supported this study.

Data availability
The data of this study is not publicly available due to its being the intellectual property of Fasa University of Medical Sciences but is available from the corresponding author on a reasonable request.

Declarations

Ethics approval and consent to participate
Approval of the research protocol: PERSIAN Cohort Study was approved by the ethics committees of the Ministry of Health and Medical Education Fasa is one of the regions. This study is in agreement with the Helsinki Declaration and Iranian national guidelines for ethics in research. (Reference number: IR.FUMS.REC.1401.007). Approval date of Registry and the Registration No.: The study was approved by Fasa University of Medical Sciences (ID: 400111, Approval Date: 2022- 4–25). Informed consent: Informed written consent was obtained from all participants.

Competing interests
The authors declare no competing interests.

Author details
[1]Social Determinants in Health Promotion Research Center, Hormozgan Health Institute, Hormozgan University of Medical Sciences, Bandar Abbas, Iran
[2]Student Research Committee, Shiraz University of Medical Sciences, Shiraz, Iran
[3]Department of Clinical Nutrition, School of Nutrition and Food Sciences, Shiraz University of Medical Sciences, Shiraz, Iran
[4]Department of Biostatistics and Epidemiology, Faculty of Health and Nutrition, Bushehr University of Medical Sciences, Bushehr, Iran
[5]The Persian Gulf Tropical Medicine Research Center, The Persian Gulf Biomedical Sciences Research Institute, Bushehr University of Medical Sciences, Bushehr, Iran
[6]Social Determinants in Health Promotion Research Center, Hormozgan Health Institute, Hormozgan University of Medical Sciences, Bandar Abbas, Iran
[7]Basic and Molecular Epidemiology of Gastrointestinal Disorders Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[8] Non-communicable disease research center, Fasa University of Medical Sciences, Fasa, Iran

Received: 2 December 2023 / Accepted: 23 August 2024
Published online: 03 September 2024

References

1. Ramezankhani A, Harati H, Bozorgmanesh M, Tohidi M, Khalili D, Azizi F et al. Diabetes mellitus: findings from 20 years of the Tehran lipid and glucose study. Int J Endocrinol Metabolism. 2018;16(4 Suppl).
2. Wang X, Zhai M, Ren Z, Ren H, Li M, Quan D, et al. Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. BMC Med Inf Decis Mak. 2021;21(1):1–14.
3. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z. Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. BMJ Nutrition, Prevention & Health. 2021:bmjnph-2020-000200.
4. Terefe AN, Gelaw AB. Modeling time-to-recovery of Adult Diabetic patients using Cox-Proportional hazards Model. Int J Stat Distrib Appl. 2017;3(4):67.
5. Mirzaei M, Rahmaninan M, Mirzaei M, Nadjarzadeh A. Epidemiology of diabetes mellitus, pre-diabetes, undiagnosed and uncontrolled diabetes in Central Iran: results from Yazd health study. BMC Public Health. 2020;20(1):1–9.
6. Fonseca VA. Defining and characterizing the progression of type 2 diabetes. Diabetes Care. 2009;32(suppl 2):S151–6.
7. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes mellitus with machine learning techniques. Front Genet. 2018;9:515.
8. Johnson M, Fishbein H, Jeffries Birch R, Yu Q, Mardon R et al. HbA1c Evidence for a Prediabetes Diagnosis Delays Onset of Type 2 Diabetes. J Endocrinol Sci. 2021;3(1).
9. Yosefian I, Mosa Farkhani E, Baneshi MR. Application of random forest survival models to increase generalizability of decision trees: a case study in acute myocardial infarction. Computational and mathematical methods in medicine. 2015;2015.
10. Sattar A, Argyropoulos C, Weissfeld L, Younas N, Fried L, Kellum JA, et al. All-cause and cause-specific mortality associated with diabetes in prevalent hemodialysis patients. BMC Nephrol. 2012;13(1):1–9.
11. Karimi N, Safari M, Mirzaei M, Kassaeian A, Roshanaei G, Omidi T. Determining the factors affecting the survival of HIV patients: comparison of Cox Model and the Random Survival Forest Method. Disease Diagnosis. 2019;8(2):124–9.
12. Cetin S, Ulgen A, Dede I, Li W. On Fair Performance comparison between Random Survival Forest and Cox Regression: an example of Colorectal Cancer Study. SciMedicine J. 2021;3(1):66–76.
13. Tran TT, Lee J, Gunathilake M, Kim J, Kim S-Y, Cho H, et al. A comparison of machine learning models and Cox proportional hazards models regarding their ability to predict the risk of gastrointestinal cancer based on metabolic syndrome and its components. Front Oncol. 2023;13:1049787.
14. Aivaliotis G, Palczewski J, Atkinson R, Cade JE, Morris MA. A comparison of time to event analysis methods, using weight status and breast cancer as a case study. Sci Rep. 2021;11(1):14058.
15. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep. 2020;10(1):1–12.
16. Dietrich S, Floegel A, Troll M, Kühn T, Rathmann W, Peters A, et al. Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. Int J Epidemiol. 2016;45(5):1406–20.
17. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT. Application of random forests methods to diabetic retinopathy classification analyses. PLoS ONE. 2014;9(6):e98587.
18. Farjam M, Bahrami H, Bahramali E, Jamshidi J, Askari A, Zakeri H, et al. A cohort study protocol to analyze the predisposing factors to common

chronic non-communicable diseases in rural areas: Fasa Cohort Study. BMC Public Health. 2016;16(1):1–8.
19. Bansal N. Prediabetes diagnosis and treatment: a review. World J Diabetes. 2015;6(2):296.
20. Sung K-C, Wild SH, Byrne CD. Resolution of fatty liver and risk of incident diabetes. J Clin Endocrinol Metabolism. 2013;98(9):3637–43.
21. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2014;37(Supplement1):S81–90.
22. George B, Seals S, Aban I. Survival analysis and regression models. J Nuclear Cardiol. 2014;21(4):686–94.
23. Qiu X, Gao J, Yang J, Hu J, Hu W, Kong L, et al. A comparison study of machine learning (random survival forest) and classic statistic (cox proportional hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. Front Oncol. 2020;10:551420.
24. Jung J-O, Crnovrsanin N, Wirsik NM, Nienhüser H, Peters L, Popp F, et al. Machine learning for optimized individual survival prediction in resectable upper gastrointestinal cancer. J Cancer Res Clin Oncol. 2023;149(5):1691–702.
25. Alabdallah A, Ohlsson M, Pashami S, Rögnvaldsson T. The Concordance Index decomposition–A measure for a deeper understanding of survival prediction models. arXiv Preprint arXiv:220300144. 2022.
26. Man RE, Charumathi S, Gan ATL, Fenwick EK, Tey CS, Chua J, et al. Cumulative incidence and risk factors of prediabetes and type 2 diabetes in a Singaporean malay cohort. Diabetes Res Clin Pract. 2017;127:163–71.
27. Mohan V, Deepa M, Anjana R, Lanthorn H, Deepa R. Incidence of diabetes and pre-diabetes in a selected urban south Indian population (CUPS-19). J Assoc Physicians India. 2008;56:152–7.
28. Goñi MJ, Forga L, Ibañez B, Cambra K, Mozas D, Anda E. Incidence and risk factors involved in the development of nephropathy in patients with type 1 diabetes mellitus: follow up since onset. Can J Diabetes. 2016;40(3):258–63.
29. Kautzky-Willer A, Stich K, Hintersteiner J, Kautzky A, Kamyar MR, Saukel J, et al. Sex-specific-differences in cardiometabolic risk in type 1 diabetes: a cross-sectional study. Cardiovasc Diabetol. 2013;12(1):1–11.
30. Tonolo G. Sex-gender awareness in diabetes. Diabetology. 2021;2(2):117–22.
31. Arnetz L, Ekberg NR, Alvarsson M. Sex differences in type 2 diabetes: focus on disease course and outcomes. Diabetes Metabolic Syndrome Obesity: Targets Therapy. 2014;7:409.
32. Li T, Quan H, Zhang H, Lin L, Lin L, Ou Q, et al. Type 2 diabetes is more predictable in women than men by multiple anthropometric and biochemical measures. Sci Rep. 2021;11(1):1–10.
33. Jeon J, Jung KJ, Jee SH. Waist circumference trajectories and risk of type 2 diabetes mellitus in Korean population: the Korean genome and epidemiology study (KoGES). BMC Public Health. 2019;19(1):1–11.
34. Tatsumi Y, Watanabe M, Nakai M, Kokubo Y, Higashiyama A, Nishimura K et al. Changes in waist circumference and the incidence of type 2 diabetes in community-dwelling men and women: the Suita Study. J Epidemiol. 2015:JE20140160.
35. McLaughlin T, Lamendola C, Liu A, Abbasi F. Preferential fat deposition in subcutaneous versus visceral depots is associated with insulin sensitivity. J Clin Endocrinol Metabolism. 2011;96(11):E1756–60.
36. Wannamethee SG, Shaper AG. Weight change and duration of overweight and obesity in the incidence of type 2 diabetes. Diabetes Care. 1999;22(8):1266–72.
37. Fan Y, Wang R, Ding L, Meng Z, Zhang Q, Shen Y, et al. Waist circumference and its changes are more strongly associated with the risk of type 2 diabetes than body mass index and changes in body weight in Chinese adults. J Nutr. 2020;150(5):1259–65.
38. Lee J, Cho YK, Kang YM, Kim HS, Jung CH, Kim H-K, et al. The impact of NAFLD and waist circumference changes on diabetes development in prediabetes subjects. Sci Rep. 2019;9(1):1–8.
39. Graham EA, Deschenes SS, Khalil MN, Danna S, Filion KB, Schmitz N. Measures of depression and risk of type 2 diabetes: a systematic review and meta-analysis. J Affect Disord. 2020;265:224–32.
40. Nouwen A, Adriaanse M, van Dam K, Iversen MM, Viechtbauer W, Peyrot M, et al. Longitudinal associations between depression and diabetes complications: a systematic review and meta-analysis. Diabet Med. 2019;36(12):1562–72.
41. Meex RC, Watt MJ. Hepatokines: linking nonalcoholic fatty liver disease and insulin resistance. Nat Reviews Endocrinol. 2017;13(9):509–20.
42. Tilg H, Moschen AR, Roden M. NAFLD and Diabetes Mellitus. Nat Reviews Gastroenterol Hepatol. 2017;14(1):32–42.
43. Yamazaki H, Tsuboya T, Tsuji K, Dohke M, Maguchi H. Independent association between improvement of nonalcoholic fatty liver disease and reduced incidence of type 2 diabetes. Diabetes Care. 2015;38(9):1673–9.
44. Guo VY, Cao B, Wong CKH, Yu EYT. The association between daytime napping and risk of diabetes: a systematic review and meta-analysis of observational studies. Sleep Med. 2017;37:105–12.
45. Zhang S, Xie L, Yu H, Zhang W, Qian B. Association between nighttime-daytime sleep patterns and chronic diseases in Chinese elderly population: a community-based cross-sectional study. BMC Geriatr. 2019;19(1):1–10.
46. Mason IC, Qian J, Adler GK, Scheer FA. Impact of circadian disruption on glucose metabolism: implications for type 2 diabetes. Diabetologia. 2020;63(3):462–72.
47. Masa JF, Rubio M, Pérez P, Mota M, Sánchez de Cos J, Montserrat JM. Association between habitual naps and sleep apnea. Sleep. 2006;29(11):1463–8.
48. Tasali E, Mokhlesi B, Van Cauter E. Obstructive sleep apnea and type 2 diabetes: interacting epidemics. Chest. 2008;133(2):496–506.
49. Rajan P, Greenberg H. Obstructive sleep apnea as a risk factor for type 2 diabetes mellitus. Nat Sci Sleep. 2015;7:113.
50. Beulens J, Rutters F, Ryden L, Schnell O, Mellbin L, Hart H, et al. Risk and management of pre-diabetes. Eur J Prev Cardiol. 2019;26(2suppl):47–54.
51. Kebbe M, Sparks JR, Flanagan EW, Redman LM. Beyond weight loss: current perspectives on the impact of calorie restriction on healthspan and lifespan. Expert Rev Endocrinol Metabolism. 2021;16(3):95–108.
52. Mundi MS, Velapati S, Patel J, Kellogg TA, Abu Dayyeh BK, Hurt RT. Evolution of NAFLD and its management. Nutr Clin Pract. 2020;35(1):72–84.
53. An Q, Yu Z, Sun F, Chen J, Zhang A. The effectiveness of cognitive behavioral therapy for Depression among individuals with diabetes: a systematic review and Meta-analysis. Curr Diab Rep. 2023;23(9):245–52.
54. Safari M, Abbasi M, Gohari Ensaf F, Berangi Z, Roshanaei G. Identification of factors affecting metastatic gastric cancer patients' survival using the random survival forest and comparison with cox regression model. Iran J Epidemiol. 2020;15(4):343–51.
55. Morsy S, Hieu TH, Makram AM, Hassan OG, Duc NTM, Zayan A et al. Is it time to use machine learning survival algorithms for survival and risk factors prediction instead of Cox proportional hazard regression? A comparative population-based study. medRxiv. 2021.
56. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18:1–12.
57. Krzyziński M, Spytek M, Baniecki H, Biecek P. SurvSHAP (t): time-dependent explanations of machine learning survival models. Knowl Based Syst. 2023;262:110234.

Publisher's note