Construction of risk prediction model of type 2 diabetes mellitus based on logistic regression

Jian Li[1,#], Qin Huang[2], Minghua Dong[3,#] , Wei Qiu[3], Lixia Jiang[4], Xiaoting Luo[5], Zhengchun Huang[1], Shuiqin Chen[5], Qinfeng Wu[3], Lu Ou-Yang[3], Qin Wu[4], Lihua Liu[5] and *Shumei Li*[3, a]
[1]*Department of Anatomy Gannan Medical College, Ganzhou, Jiangxi, China*
[2]*Department of General Medicine, St.Michael hospital, Shanghai, China*
[3]*Department of Epidemiology in Preventive Medicine, Gannan Medical College, Ganzhou, Jiangxi, China*
[4]*Department of Clinical laboratory of 1'th affiliate hospital, Gannan Medical College, Ganzhou, Jiangxi, China*
[5]*Department of Biochemistry and Molecular Biology, Gannan Medical College, Ganzhou, Jiangxi, China*

Abstract. Objective: to construct multi factor prediction model for the individual risk of T2DM, and to explore new ideas for early warning, prevention and personalized health services for T2DM. Methods: using logistic regression techniques to screen the risk factors for T2DM and construct the risk prediction model of T2DM. Results: Male's risk prediction model logistic regression equation: logit(P)=BMI × 0.735+ vegetables × (-0.671) + age × 0.838+ diastolic pressure × 0.296+ physical activity× (-2.287) + sleep ×(-0.009) +smoking ×0.214; Female's risk prediction model logistic regression equation: logit(P)=BMI ×1.979+ vegetables× (-0.292) + age × 1.355+ diastolic pressure× 0.522+ physical activity × (-2.287) + sleep × (-0.010).The area under the ROC curve of male was 0.83, the sensitivity was 0.72, the specificity was 0.86, the area under the ROC curve of female was 0.84, the sensitivity was 0.75, the specificity was 0.90. Conclusion: This study model data is from a compared study of nested case, the risk prediction model has been established by using the more mature logistic regression techniques, and the model is higher predictive sensitivity, specificity and stability.

1 Introduction

With the rapid development of social economy, people's life style and dietary structure have changed and population aging intensified, diabetes prevalence rate rose rapidly, especially in our country [1], it has brought heavy economic burden to society and family. In recent years, foreign countries have tended to use a risk assessment tool to the risk of type 2 diabetes mellitus(T2DM) for prediction and risk score, for finding the early identification of high-risk groups, to control risk factors by carrying out the health education and lifestyle intervention , and then reduce the T2DM [2]. Some domestic scholars also constructed T2DM risk model with risk factors scoring method and OR data score and so on, to predict high-risk T2DM;and, in this aspect, they did some propaganda and education to make people accept the T2DM risk individual prediction model [3-4] in the identification of high-risk

[a] Corresponding author: Shumei Li, gnyxylsm@163.com. Phone: 008615083787928.
[#]These authors contributed equally to this study and share first authorship.
This study was supported by scientific fund from National Natural Science Fund in China (No.81360445), Science and technology support program of Jiangxi Province (NO.20132BBG70086).

groups gradually .

2 Research design and program

2.1 Research objects

Researchers chose the sample from T2DM baseline survey for the project study in 19 districts and counties in Ganzhou City in 2009.The study subject is a total of 8086 copies of data, age 35-64 years old. The baseline survey and follow-up data content include general demographic characteristics, smoking and drinking history, personal health history, diabetes family history, physical activity and physical exercise and so on; human body measurements included height, weight, BMI, waist circumference, hip circumference, waist hip ratio (WHR), blood pressure, lung capacity and so on ; biochemical metabolic index included fasting blood glucose (FBG), total cholesterol (TC), triglyceride (TG), high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), blood uric acid (UA) (save all to MSSQL database).

2.2 Research methods

Researchers randomly selected 2 / 3 samples (5391persons) from MSSQL database as the training group, the pre-selected variables is age, sex, BMI, waist circumference and waist to hip ratio (WHR), blood pressure (BP), T2DM family history, physical activity, smoking and drinking, FBG, TC, TG, HDL-C, LDL-C; and then made single factor analysis for the pre-seleced variables . Multivariate analysis mainly adopted mature logistic regression model, as to whether diabetes as the dependent variable (diabetic patient with the value 1, non diabetic patient with the value 0), other variables (suspected risk factor variables) as independent variables were multivariate logistic regression, logistic(p)= $\ln(\frac{p}{1-p})$,to construct the T2DM risk prediction model. In order to increase the prediction accuracy, prediction model of gender were fitted.Another 1/3 sample (2695 people) in the study cohort was used as the test group, the area under the receiver operating characteristic curve (ROC) was used to analyze the sensitivity and specificity of the model, to evaluate the predictive effect of the model .

3 Results

3.1 Relationship between age, sex and T2DM

Research data showed the prevalence of T2DM was 7.7%, with group of every 5 years , people will be divided into 35-, 40-, 45-, 50-, 55-, 60-64, T2DM prevalence was compared between six groups by gender , the differences of T2DM prevalence between each group were not statistically significant ($p > 0.05$).(see Table 1).

Table 1. Different gender and age in T2DM prevalence rate.

age(year)	male			female		
	n	T2DM	%	n	T2DM	%
35 -	288	8	2.8	368	13	3.5
40-	362	20	5.5	439	12	2.7
45-	375	22	5.9	402	21	5.2
50-	273	22	8.1	395	28	7.1
55-	390	50	12.8	475	41	8.6
60-64	378	53	14.0	363	57	15.7
total	2066	175	8.5	2442	172	7.0

3.2 Relationship between the body mass index (kg / M^2) and T2DM

Body mass index (BMI) were divided into three groups: the normal group (18.5-23.9),T2DM prevalence was 6.7% (88/1315);hyper height group(≥24.0), T2DM prevalence was 9.6% (86/897); obesity group(≥28), T2DM prevalence was 14.9 % (35/235). The three groups were compared with the increase of BMI and the prevalence of T2DM, the difference was statistically significant (x^2=19.15, p<0.001).

BMI group by gender: Male, T2DM prevalence of normal group was 7.3% (42/578);T2DM prevalence of overweight group was 9.9% (41/414); T2DM prevalence of obesity group was 17.2% (16/93); There are differences between these groups and these differences were statistical significant (x^2=10.03, p < 0.05). Female, T2DM prevalence of the normal group was 6.2% (46/737); T2DM prevalence of overweight group was 9.3% (45/483); T2DM prevalence of obesity group was 13.4% (19/142); There are differences between these groups and these differences was statistical significant (x^2=9.72, p < 0.05).(see Table 2).

Table 2. T2DM prevalence rate of different BMI In different gender.

BMI	male			female		
(kg/m^2)	n	T2DM	%	n	T2DM	%
18.5-23.9	578	42	7.3	737	46	6.2
≥24.0	414	41	9.9	483	45	9.3
≥28	93	16	17.2	142	19	13.4
total	1085	99	9.1	1362	110	8.1

3.3 Relationship between waist circumference and waist to hip ratio of T2DM

Waist samples were divided into two groups: normal group (male <85cm, female <80cm), T2DM prevalence was 5.7% (79/1377); abnormal group (male ≥85cm, the female≥80cm), T2DM prevalence was12.2% (129/1060). Waist circumference increased, the prevalence of T2DM increased, the difference was statistically significant (x^2=31.74, p<0.001).Male, T2DM prevalence of normal group was 7.0% (45/640);T2DM prevalence of abnormal group was 12.2% (54/443); There are differences between these groups and these differences were statistical significant (x^2=8.39, p < 0.05). Female, T2DM prevalence of the normal group was 4.6% (34/737); T2DM prevalence of abnormal group was 12.2% (75/617); There are differences between these groups and these differences was statistical significant (p < 0.05).(see Table 3).

Table 3. Relationship between waist circumference and T2DM

waist (cm)	male			female		
	n	T2DM	%	n	T2DM	%
male <85, female <80	640	45	7.0	737	34	4.6
male ≥85, female≥80	443	54	12.2	617	75	12.2
total	1083	99	9.1	1354	109	8.1

The samples was divided into three groups by the ratio of waist to hip(W-H), the low group (male WH<0.82, female WH<0.78), the prevalence of T2DM was 3.9% (16/410); the normal group (male 0.82≤WH< 0.91, female 0.78≤WH<0.87) ,the prevalence of T2DM was 6.6% (83/1267); the high group (male WH≥0.91, female≥0.87), the prevalence of T2DM was 14.2%, 107/754). Compared with the three groups between male and female, WH increased, the prevalence of T2DM increased, the difference was statistically significant (p<0.05).(see Table 4).

Table 4. Waist hip ratio and the prevalence of T2DM relations.

W-H	male			female		
	n	T2DM	%	n	T2DM	%
male <0.82, female <0.78	207	13	6.3	203	3	1.5
male 0.82-0.91, female 0.78-0.87	588	44	7.5	679	39	5.7
male ≥0.91, female ≥0.87	286	40	14.0	468	67	14.3
total	1081	97	9.0	1350	109	8.1

3.4 Relationship between blood pressure and the prevalence of T2DM

The samples were divided into two groups by the blood pressure,the normal group (systolic blood pressure SBP<140 mmHg and diastolic blood pressure DBP<90 mmHg), and hypertension group (SBP≥140 mmHg, DBP ≥90 mmHg) .
Normal group of Systolic blood pressure(SBP)'s T2DM prevalence was 6.9% (132/1920). Hypertension group of SBP 's T2DM prevalence was 14.8% (76/512); the prevalence of T2DM in hypertension group increased, and the difference is statistically significant (p <0.05). The normal group of diastolic blood pressure(DBP)'s T2DM prevalence was 7.6% (141/1853) ;hypertension group of DBMS's T2DM prevalence was 11.7% (67/575), the prevalence of hypertension group was different in T2DM,for male ,the difference has no statistical significance (p > 0.05);for female, the difference has statistical significance (p < 0.05).(see Table 5,6).

Table 5. SBP and T2DM prevalence rate.

SBP	male			female		
(mmHg)	n	T2DM	%	n	T2DM	%
<140	842	63	7.5	1078	69	6.4
≥140	237	35	14.8	275	41	14.9
total	1079	98	9.1	1353	110	8.1

Table 6. DBP and T2DM prevalence rate.

DBP	male			female		
(mmHg)	n	T2DM	%	n	T2DM	%
<90	805	66	8.2	1048	75	7.2
≥90	273	32	11.7	302	35	11.6
total	1078	98	9.1	1350	110	8.1

3.5 Relationship between total cholesterol (TC) and T2DM

The total cholesterol index sample was divided into three groups: the low group ,TC < 5.18 mmol / L, T2DM prevalence was 6.2% (130/2105); intermediate group ,5.18 mmol/L≤TC<6.22 mmol/L, T2DM prevalence was 10.2% (80/783); high group ,TC≥6.22 mmol/L, T2DM prevalence was 18.6% (49/264). The prevalence of hypertension group was different in T2DM:for male and female ,the difference of group has no statistical significance (p > 0.05);for whole sample the difference has statistical significance (p < 0.05).(see Table 7)

Table 7. Regardless of gender, the relationship between cholesterol and prevalence rate of T2DM.

TC	male			female		
(mmol / L)	n	T2DM	%	n	T2DM	%
< 5.18	927	74	8.0	1178	56	4.8
5.18-6.22	328	33	10.1	455	47	10.3
≥6.22	101	14	13.9	163	35	21.5
total	1356	121	8.9	1350	138	7.7

3.6 Relationship between triglyceride (TG) and T2DM

Triglyceride sample was divided into three groups: low group ,TG < 1.70 mmol / L, the prevalence of T2DM was 6.0% (132/2209); intermediate group, 1.70 mmol/L≤TG<2.26 mmol/L, the prevalence of T2DM is 9.5% (46/482); high group , TG≥2.26 mmol/L), the prevalence of T2DM was 17.9% (82/459). Compared with the three groups between male and female, TG increased , the prevalence of T2DM increased, the difference was statistically significant ($p < 0.001$).(see Table 8).

Table 8. Relationship between triglyceride index and the prevalence rate of T2DM.

TG	male			female		
(mmol / L)	n	T2DM	%	n	T2DM	%
< 1.70	918	64	7.0	1291	68	5.3
1.70-2.26	210	18	8.1	272	28	10.3
≥2.26	225	39	17.3	234	43	18.4
total	1353	121	8.9	1797	139	7.7

3.7 Relationship between smoking and T2DM

The subjects were divided into two groups, low exposure group (no smoking) and high exposure group (smoking): there was no significant difference in the prevalence of low exposure group and high exposure group (x^2=0.831, p>0.05).

3.8 The Relationship between dietary factors and T2DM

Compared the people who often eat coarse grain with the people who don't eat the coarse grain, male's OR was 0.41, 95% confidence interval (0.25, 0.67), The differences of prevalence have statistical significance (x^2=13.12. p< 0.001), coarse grain is the protective factor; Female's OR is 0.73, 95% confidence interval for (0.47, 1.50), the differences of prevalence have statistical significance (x^2=13.18, p < 0.05); coarse grains is the protective factor.

Compared the people who often edible meat with the people who don't eat meat,male's OR is 2.37, 95% confidence interval (1.46, 3.84), the differences in prevalence have statistical significance (x^2=12.92, p < 0.001), meat is the risk factor ; The difference of female prevalence have no statistical significance (x^2= 0.26, p> 0.05).

Compared the people who often edible vegetable with the people who don't edible the vegetables, male's OR is 0.59, 95% confidence interval for (0.37, 0.93), the differences in prevalence have statistical significance (x^2=5.16, p < 0.05),vegetable is the protective factor; Female's OR is 0.50, 95% confidence interval for (0.32, 0.77),the differences in prevalence have statistical significance (x^2=9.92, p < 0.05); vegetable is the protective factor.

3.9 Multiple factors analysis of logistic regression

The analysis of Significant predictive variables had showed : male subjects in the study taking BMI, age, diastolic blood pressure, smoking as the risk factor, and vegetable intake, physical activity, sleep as the protective factor (β<0) ,and risk prediction model regression equation logit (P) = BMI ×0.735+ vegetables× (-0.671) + age ×0.838+ diastolic pressure × 0.296+ physical activity×(-2.287) sleep ×(-0.009) +smoking ×0.214. The female subjects in the study taking BMI, age, diastolic pressure as the risk factor, and vegetable intake, physical activity, sleep as the protective factor, risk prediction model regression equation logit (P) = BMI ×1.979+ vegetables × (-0.292) + age× 1.355+ diastolic pressure × 0.522+ physical activity × (-2.287) + sleep × (-0.010). (see Table 9,10)

Table 9. The Results of logistic regression analysis of male relative indexes.

Variable beta	β	S.E	x^2	P
Body mass index	0.735	0.905	0.659	0.417
Vegetable	-0.671	0.292	5.291	0.021
Age	0.838	0.497	2.843	0.092
Diastolic blood pressure	0.296	0.303	0.956	0.328
Physical activity	-2.287	0.106	465.380	0.000
Sleep	-0.009	0.166	0.003	0.956
Smoking	0.214	0.235	0.828	0.363

Table 10. The Results of logistic regression analysis of female related indexes.

Variable beta	β	S.E	x^2	P
Body mass index	1.979	1.056	3.512	0.061
Vegetable	-0.292	0.316	0.855	0.355
Age	1.355	0.588	5.318	0.021
Diastolic blood pressure	0.522	0.333	2.460	0.117
Physical activity	-2.287	0.106	465.380	0.000
sleep	-0.010	0.123	0.007	0.956

3.10 The Internal authenticity of the model and its evaluation

Male' s area under the ROC curve area was 0.83, sensitivity was 0.72, specificity was 0.86; female's area under ROC curve was 0.84, sensitivity was 0.75, specificity was 0.90, two sets of data showed that model had a high predictive value.(see Figure 1).

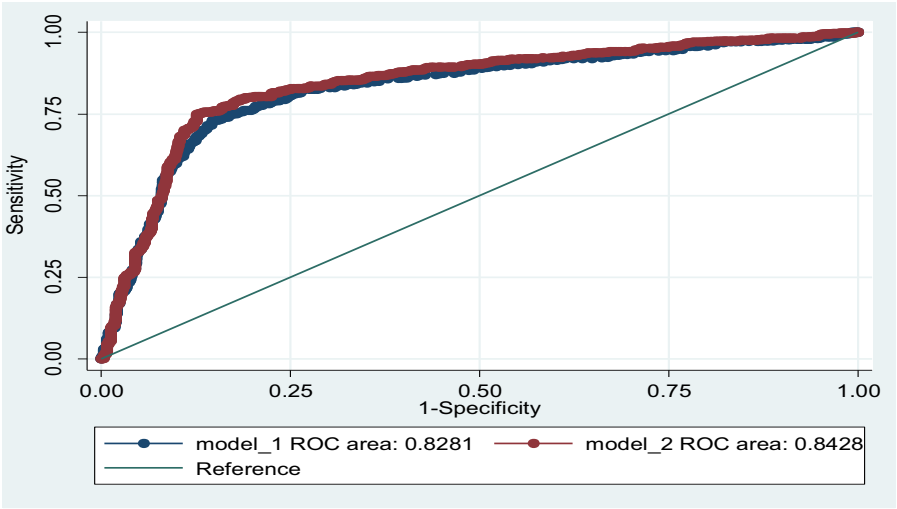

Figure 1. Sensitivity and specificity of the model.

4 Discussion

The modeling data was from compared study of prophase nested case, and used the less error and more mature logistic regression techniques in theory to deal with the variables. The risk model that had been established on this base has higher sensitivity, specificity and stability. The development of T2DM risk model started late. Since 2003, some type 2 diabetes risk assessment tools had been

gradually developed and applicated in foreign country [5-9], but due to the difference in race, behavior and lifestyle, diet, medical and health services, economic and cultural factors, foreign risk assessment tool was not fully applicable to Asian and / or Chinese [10-11].Therefore, since 2006, scholars in our country [12-14] have concluded that the main risk factors of diabetes mellitus and its relative risk to establish evaluation methods and evaluation model of adult individual diabetes risk in our country ,on the basis of epidemiological survey of diabetes ,by risk factors scoring method, disease risk score, the analysis of diabetes risk factors and prevalence data. These models mostly based on the compared study cases , prospective study was insufficient; Our study team has fully considered the conditions which required for robust risk model, include stable and representative large samples , a prospective observation, multivariate estimation and mature computer technology platform .But the research is still in initial stage with many remaining problems: the first is the study did not exclude the influence of the linear of predictive variables in model to the stability of the model; researchers only evaluate the predictive model by the similar people , part of the data is not so complete; We will evaluate and verificate this model by screening return visitors in the latter part of the study .

References

1. Yang WY, Lu JM, Weng Jp, N Engl J Med ,*Prevalence of diabetes among men and women in China*,**362**,1090-1101(2010).
2. Douglas N, Rohini M,Tom D, BMJ,*Risk models and scores for type 2 diabetes: systematic review*, **343**, d7163(2011).
3. Mann DM, Bertoni AG, Shimbo D, Am J Epidemiol, *Comparative Validity of 3 Diabetes Mellitus Risk prediction Scoring Models in a Multiethnic US Cohort The Multi-Ethnic Study of Atherosclerosis*, **171**,980–988(2010).
4. Almeda-Valdes p, Cuevas-Ramos D, Mehta R,Curr Diabetes Rev,*UKpDS Risk Engine, decode and diabetespHD models for the estimation of cardiovascular risk in patients with diabetes,* **6**,1-8(2010).
5. Buijsse B, Simmons RK, Griffin SJ, Schulze MB,Epidemiol Rev, *Risk Assessment Tools for Identifying Individuals at Risk of Developing Type 2 Diabetes,* **33**,46-62(2011).
6. Lindström J, Tuomilehto J, Diabetes Care,*The diabetes risk score: a practical tool to predict type 2 diabetes risk*,**26**,725–731(2003).
7. Schulze MB, Weikert C, pischon T, Diabetes Care,*Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EpIC-potsdam Study*,**32**,2116–2119(2009).
8. Glmer,Charlotte,Carstensen,Bendix,Sandbk,Annelli,Lauritzen,Torsten,Jrgensen,Torben,Borch-Johnsen,Knut, Diabetes Care ,*A Danish Diabetes Risk Score for Targeted Screening* ,**3**,727-733(2004).
9. Heikes KE, Eddy DM, Arondekar B, Diabetes Care, *Diabetes Risk Calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes*,**31**,1040-1045(2008).
10. Liu M,Pan CY, Jing MM, Su HY,Lu JM, Chinese Journal of diabetes, *Evaluation of diabetes risk score in the evaluation of new onset diabetes*,**17**,201 -204(2009).
11. 12.Calvin Woon-Loong Chin, Elia n Hui San Chia, Stefan Ma, Derrick Heng, Maud rene Tan, Jean ette Lee, E Shyong Tai, Agus Salim, BMC,*The ARIC predictive model reliably predicted risk of type II diabetes in Asian populations*,**1** -13(2012).
12. Li XY, Li R, Zhang SN, Chinese public health, *Effect evaluation of different screening methods for asymptomatic diabetes*,**22**,687-689(2006).
13. Wu HY, Pan P, He Y, Chinese Journal of health management, *The risk assessment method for adults with diabetes in China*,**1**,95-98(2007).
14. Zhang L, Shi K, Yi D, Chongqing medical, *Study on risk assessment of diabetes in community residents in Chongqing City*,**40**,1885-1888(2011).