Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Predictive modeling the probability of suffering from metabolic syndrome using machine learning: A population-based study

Xiang Hu[a,b,1], Xue-Ke Li[a,b,1], Shiping Wen[c], Xingyu Li[d], Tian-Shu Zeng[a,b], Jiao-Yue Zhang[a,b], Weiqing Wang[e], Yufang Bi[e], Qiao Zhang[f], Sheng-Hua Tian[a,b], Jie Min[a,b], Ying Wang[a,b], Geng Liu[a,b], Hantao Huang[g], Miaomiao Peng[a,b], Jun Zhang[g], Chaodong Wu[h], Yu-Ming Li[a,b], Hui Sun[a,b], Guang Ning[e], Lu-Lu Chen[a,b,*]

[a] *Department of Endocrinology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China*
[b] *Hubei Provincial Clinical Research Center for Diabetes and Metabolic Disorders, Wuhan, China*
[c] *Centre for Artificial Intelligence, Faculty of Engineering Information Technology, University of Technology Sydney, Ultimo, NSW, 2007, Australia*
[d] *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China*
[e] *Department of Endocrinology and Metabolism, State Key Laboratory of Medical Genomes, National Clinical Research Center for Metabolic Diseases, Shanghai Clinical Center for Endocrine and Metabolic Diseases, Shanghai Institute of Endocrine and Metabolic Diseases, Ruijin Hospital, Shanghai Jiao-Tong University School of Medicine, Shanghai, China*
[f] *Department of Cardiovascular Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China*
[g] *Yiling Hospital, Yichang, China*
[h] *Department of Nutrition and Food Science, Texas A&M University, College Station, TX, USA*

## ARTICLE INFO

*Keywords:*
Metabolic syndrome
Machine learning
Prediction model
Predictors
Population-based study

## ABSTRACT

*Background:* There is an increasing trend of Metabolic syndrome (MetS) prevalence, which has been considered as an important contributor for cardiovascular disease (CVD), cancers and diabetes. However, there is often a long asymptomatic phase of MetS, resulting in not diagnosed and intervened so timely as needed. It would be very helpful to explore tools to predict the probability of suffering from MetS in daily life or routinely clinical practice.
*Objective:* To develop models that predict individuals' probability of suffering from MetS timely with high efficacy in general population.
*Methods:* The present study enrolled 8964 individuals aged 40–75 years without severe diseases, which was a part of the REACTION study from October 2011 to February 2012. We developed three prediction models for different scenarios in hospital (Model 1, 2) or at home (Model 3) based on LightGBM (LGBM) technique and corresponding logistic regression (LR) models were also constructed for comparison. Model 1 included variables of laboratory tests, lifestyles and anthropometric measurements while model 2 was built with components of MetS excluded based on model 1, and model 3 was constructed with blood biochemical indexes removed based on model 2. Additionally, we also investigated the strength of association between the predictive factors and MetS, as well as that between the predictors and each component of MetS.
*Results:* In this study, 2714 (30.3%) participants suffer from MetS accordingly. The performances of the LGBM models in predicting the probability of suffering from MetS produced good results and were presented as follows: model 1 had an area under the curve (AUC) value of 0.993 while model 2 indicated an AUC value of 0.885. Model 3 had an AUC value of 0.859, which is close to that of model 2. The AUC values of LR model 1 and 2 for the scenario in hospital and model 3 at home were 0.938, 0.839 and 0.820 respectively, which seemed lower than that of their corresponding machine learning models, respectively. In both LGBM and logistic models, gender, height and resting pulse rate (RPR) were predictors for MetS. Women had higher risk of MetS than men (OR 8.84, CI: 6.70–11.66), and each 1-cm increase in height indicated 3.8% higher risk of suffering from MetS in people over 58 years, whereas each 1- Beat Per Minute (bpm) increase in RPR showed 1.0% higher risk in individuals younger than 62 years.

[*] Corresponding author.
*E-mail address:* cheria_chen@126.com (L.-L. Chen).

[1] X.H. and X.-K. L. are joint first authors, each contributing equally.

https://doi.org/10.1016/j.heliyon.2022.e12343
Received 13 February 2022; Received in revised form 16 June 2022; Accepted 6 December 2022

*Conclusion:* The present study showed that the prediction models developed by machine learning demonstrated effective in evaluating the probability of suffering from MetS, and presented prominent predicting efficacies and accuracies. Additionally, we found that women showed a higher risk of MetS than men, and height in individuals over 58 years was important factor in predicting the probability of suffering from MetS while RPR was of vital importance in people aged 40–62 years.

## 1. Introduction

Metabolic syndrome (MetS) is a constellation of multiple metabolic abnormalities, including glucose intolerance, central obesity, dyslipidemia, and hypertension [1]. These conditions may group in one individual simultaneously, which are the causative factors for diabetes and cardio-cerebrovascular diseases. MetS is often associated with other health problems, such as fatty liver, cholesterol gallstones, obstructive sleep apnea, gout, depression, musculoskeletal disease, and polycystic ovarian syndrome [2]. Notably, the prevalence of MetS is rapidly increasing worldwide, affecting more than 20% of the population in the USA, China and Europe [3, 4, 5], and imposing a substantial health economic burden on the world. It is reported MetS is associated with a 2-fold increase in the risk of cardio-cerebrovascular disease, and a 1.5-fold increase in the risk of all-cause mortality [6]. Therefore, it is increasingly considered that it is imperative to pay more attention to improving preventive and therapeutic strategies to achieve a better control. Noteworthy, there is often a long asymptomatic phase, which easily make MetS miss diagnosis. Moreover, a large number of people do not know whether they are susceptible to MetS and its subsequent complications. Therefore, it is vital for medical practitioners and individuals to timely and accurately assess the probability of suffering from MetS, which is important for early diagnosis and intervention of these high-probability people, as well as alleviation of their probable health and economic burden.

It is increasingly considered that the etiology of MetS is complex and a series of risk factors were involved such as rapidly changes in lifestyle, socio-economics, races [7, 8, 9], sex- and age-related determinants [10, 11] which are concluded by using logistic regression methods previously. However, there might be more potential factors associating with MetS. Additionally, it seems a little difficult to filter a large number of predictive factors from numerous factors and simultaneously evaluate their strength of association by traditional statistical approaches [12]. It is reported that the predictive models established by traditional statistical methods (such as logistic regression, Cox regression) and shallow machine learning algorithms (Shallow Learning) have great accuracy [13]. There are still some important concerns of the multicollinearity and interaction of variables, which may affect the predictive ability of the model. Moreover, the internal correlation of repeated observation data for the same individual is somewhat difficult to separate by classical methods, and the information of high-dimensional features might not to be mined and exploited effectively enough [14]. Recently, artificial intelligence (AI) is widely utilized due to the explosively increasing amount of data and the effective performance of intelligent technologies to handle the massive information [15]. There are an increasing number of studies using machine learning applied to medical imaging to assist in diagnosis, such as pattern recognition which has been used in clinical practice [16, 17, 18]. Numerous studies indicated that machine learning method has an efficient and impressive predictive power using gradient boosting technique to increase robustness and reduce variance of predictions [19, 20]. Among multitudinous techniques, LightGBM (LGBM) is new Gradient Boosting Decision Tree (GBDT) implementation, which is a popular machine learning algorithm and speeds up the training process of conventional GBDT by up to over 20 times while achieving almost the same accuracy [21] and has a great advantage in dealing with large number of data instances and large number of features, which is commonly involved in MetS.

Thus, we test to use LGBM to develop models to predict the probability of suffering from MetS in different scenarios (in daily life at home or in clinical practice in hospital) and investigate the new probable predicators.

## 2. Materials and methods

### 2.1. Study design and population

The present study was part of the Risk Evaluation of cAncers in Chinese diabeTic Individuals: a lONgitudinal (REACTION) population-based cohort study performed from October 2011 and February 2012 in Yi Chang reported previously [22]. In this study, aged 40 and 75 years were enrolled in view that people aged over 40 years are at higher risk of developing metabolic syndrome [23], and the risk of participating in the investigation might be a somewhat higher in people over 75 years old, which is very close to the average life expectancy of Chinese people [24]. Data were taken by trained staff using a standard questionnaire and clinical measurements in the selected communities. The study was approved by the Ethics Committee of Huazhong University of Science and Technology, and all participants gave signed informed consent.

A flow chart illustrated the inclusion and exclusion of study subjects (Supplementary Figure 1). Overall, the original sample comprises 10186 participants. We excluded 687 participants who had severe illness (tumor, myocardial infarction, cardiovascular disease and stroke) or were unable to attend questionnaire survey and physical examination, and 525 participants for lacking anthropometric data, and 8964 individuals were includes in the final analytical sample.

### 2.2. Questionnaire, physical examination, laboratory results and diagnostic criteria

A questionnaire including demographic information, behavioral factors, personal medical history, and living habits was administered by trained staff, and physical examination and laboratory investigation were performed are described previously [25]. Height-to-weight ratio (H/W) and the homeostasis model assessment of insulin resistance (HOMA-IR) were calculated as described previously [26] and the natural logarithm of HOMA-IR (lnHOMA-IR) was employed for later analysis. The features in the original questionnaire were selected based on the risk factors of MetS investigated by the convincing literatures and some accessible and potential indicators suggested by clinical practice [27, 28].

According to the revised National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (NCEP-ATP III 2005) for Asians [2]. MetS was defined as the presence of three or more of the following five criteria proposed: 1) Central obesity: waist circumference (WC) $\geq$ 90 cm in men or WC $\geq$ 80 cm in women; 2) Hypertriglyceridemia: triglyceride (TG) $\geq$ 1.70 mmol/L; 3) Low high density lipoprotein-cholesterol (HDL-C): HDL $<$1.03 mmol/L in men or 1.29 mmol/L in women; 4) High blood pressure (BP): systolic and (or) diastolic blood pressure (SBP/DBP) $\geq$ 130/85 mmHg or taking anti-hypertensive drugs; 5) High fasting plasma glucose (FPG): FPG $\geq$5.6 mmol/L.

### 2.3. Statistical analysis

We used *t*-test for continuous variables with normal distributions and Mann-Whitney test for continuous variables with skewed distributions. Chi-squared test or Fisher exact test was employed with expects less than

10 for categorical variables to compare demographic characteristics and variables between groups.

Our primary analysis constructed different prediction models and generated a variable importance plot to assess the relative contribution of each variable for the prediction of MetS using LGBM techniques. Among the participants, 7000 (78.1%) were allocated to the training set, and 1964 (21.9%) were allocated to the test set. With the emergence of large data, LGBM trains them in sequence [29]. In each iteration, LGBM learns the decision trees by fitting the negative gradient [21]. We compared the predictive discrimination of models using the receiver operating characteristic (ROC) curve, and applied AUC to assess how well a model predicts the progression of MetS. Complementary machine learning models were developed separating men and women to account for different relative importance of predictors by gender. Furthermore, we used machine learning to investigate the strength of association between predictors and the components of MetS.

Additionally, we used a stepwise algorithm to select variables automatically for a stepwise logistic regression model to predict the probability of suffering from MetS. The data set was divided into training set and a test set (partitioned 75/25 percent). Categorical variables were excluded from the model fitting when collinearity was detected using a variable inflation factor (VIF) that was >10 [30]. Features selection process was implemented using backward elimination to rank the variables by weight starting from the model involving all variables. This process was performed to get the best simplified model by removing one variable at a time, while calculating the accuracy of the model after each removal until the accuracy's change was<0.01 [31]. The model with the lowest value of Akaike information criterion (AIC) is preferred, which assesses goodness-of-fit of the model. Predictors of MetS were estimated with adjusted odds ratios (ORs) with 95% confidence intervals (CIs) using backwards stepwise logistic regression.

The set of all variables extracted includes questionnaire responses, demographic information and laboratory results (Supplementary Table 1). The numbers of variables with uncertain or missing values for smoking status, drinking status, weight change, snore, milk-drinking, tea-drinking in the past year, culture, occupation were 306 (3.4%), 1179 (13.2%), 498 (5.6%), and 3670 (40.9%), 1135 (12.7%), 634 (7.1%), 344 (3.8%) and 446 (5.0%) individually. To reduce bias and increase statistical efficacy, we imputed missing data with multiple imputation by chained equations (MICE) [32] based on 5 replications in the R MI packages. We combined multiple analyses' results by Rubin's Rules [33] and performed sensitivity analyses using complete case for comparison (Supplementary Table 2). Since the results were similar before and after the imputation, all analyses were performed using the imputed data (Supplementary Table 3).

Data analyses were conducted with SPSS (version 22.0) software, R (version 3.4.3) and EmpowerStats (R) (www.empowerstats.com, X&Y solutions, Inc., Boston, MA). P < 0.05 was considered statistically significant.

## 3. Results

### 3.1. Study population characteristics

2714 (30.3%) individuals in total population were identified as suffering from MetS, 672 (21.2%) and 2042 (35.3%) in male and female, respectively. 1052 (11.7%) did not have any component of MetS. 3564 (39.8%) were overweight (BMI≥24 kg/m$^2$). 576 (6.4%) had history of diabetes. 1047 (64.8%) (95% confidence interval [CI] 62.4–67.2%) had newly detected diseases in 1616 diabetics group. 625 (6.9%) had history of hypertension. In a group of 5830 hypertensive patients, 5195 (89.1%) (95% CI 88.3–89.9%) had newly detected hypertension. And 116 (1.3%) had history of hyperlipidemia (Table 1). 2583 (95.7%) (95% CI 94.9–96.4%) were undetected in all 2699 participants with hyperlipidemia.

In males, the most commonly observed MetS component was elevated BP levels (67.5%), followed by elevated FPG level (59.1%) and high TG (24.5%), while elevated BP level (63.7%) was also the most common component in females, followed by elevated FPG levels (58.2%) and high WC (42.4%) (Supplementary Table 4).

### 3.2. Development and comparison of prediction models for the scenarios in hospital and at home

We imputed all 38 variables (Supplementary Table 1) into the LGBM machine learning technique to develop model 1 for the application in hospital (Figure 1A). Variables which were components for the diagnosis of MetS (TG, FPG, WC, SBP, DBP, HDL-C) were removed from model 1 to generate model 2 (Figure 1B) and to investigate the predicting efficacy of the rest factors. Considering that it was inconvenient for most people to get clinical test in daily life, we developed a prediction model 3 (home model) without blood biochemical indexes such as 2h-plasma glucose (P2hPG), lnHOMA-IR, FIns, TC, LDL-C, serum creatinine (Scr), glutamate transaminase (ALT), aspartate transaminase (AST), gamma-glutamyl transpeptidase (GGT) for people who were unlikely to participate in a health check-up timely (Figure 1C). Model 1 showed the highest AUC value (AUC = 0.993) among these three models, reflecting the best discriminative ability, and no obvious difference was observed in male and female (Table 2). The predicting efficacy of model 2 (AUC = 0.885) decreased a lot after removing the diagnostic variables of MetS. Remarkably, the predicting efficacy of model 3 was close to model 2 (AUC = 0.859) after blood biochemical indexes removed (Figure 2).

### 3.3. Predicting efficacy of the predictors for MetS, for components of MetS and in gender-separate models

The predicting efficacy of variables, which was investigated by LGBM machine learning technique, indicated that the six variables with highest importance values were all components of MetS which accorded with the diagnostic standard (Figure 1A). Our results also demonstrated that gender had high importance value in prediction model 1. The present study illustrated that factors related to dysglycemia (e.g. lnHOMA-IR, FIns and P2hPG) and obesity (waist-to-hip ratio (WHR) and waist-to-height ratio (WHTR)) had high predicting potency in model 2 (Figure 1B). In model 3 (Figure 1C), the results indicated that sex, height and RPR were important predictive factors.

Our results analyzed by machine learning indicated that lnHOMA-IR and FIns were the most important predictors for the MetS components of low HDL-C, high TG, high FPG and high WC after removing the variables directly associated with each component, respectively. Additionally, age and FPG was the best predictor for high SBP (Supplementary Figure 2).

Considering it is illustrated that gender had a high important value in predicting the probability of MetS in all models, we attempted to investigate whether the efficacy of the predictors differed in models developed by machine learning with male or female only. Our results indicated that the AUC of these gender-separate models were both lower compared with the models with the total population (Table 2). TG, FPG, SBP, DBP remained predictors with high efficacy in these gender-separate models. Remarkably, WHtR seemed to be a predictor with higher efficacy than WC in the models with male only (Supplementary Figure 3A). After removing the variables of MetS components, the predicting efficacy of WHtR, lnHOMA-IR, P2hPG and TC remained high in these gender-separate models (Supplementary Figure 3B), and these predictors had similar priority of predicting efficacy in these gender-separate models to those with total population (Supplementary Figure 3C).

### 3.4. Logistic regression prediction model for MetS

The current study also built three logistic regression (LR) models corresponding to the three machine learning models of model 1, model 2 and model as described above. The AUC values of LR model 1, 2 and 3 were 0.938, 0.839, 0.820, respectively (Table 3). The efficacy was much

**Table 1.** Baseline characteristics of study participants.

| | Male (n = 3177) | | Female (n = 5787) | | |
|---|---|---|---|---|---|
| | Normal (n = 2505) | MetS (n = 672) | Normal (n = 3745) | MetS (n = 2042) | *P* value for sex difference |
| Demographics: | | | | | |
| Age, years | 55.8 (8.6) | 54.7 (8.1) ‡ | 52.8 (8.2) | 55.7 (7.7) * | <0.001 |
| Weight, kg | 59.3 (8.6) | 70.2 (9.7) † | 52.9 (7.7) | 60.0 (8.3) * | <0.001 |
| Height, cm | 162.7 (6.1) | 164.3 (6.2) † | 153.2 (5.5) | 153.7 (5.5) ** | <0.001 |
| WC, cm | 76.9 (8.1) | 88.8 (8.0) † | 74.5 (7.7) | 84.2 (7.6) * | <0.001 |
| HC, cm | 88.9 (5.9) | 94.8 (6.4) † | 89.2 (6.2) | 94.6 (6.7) * | <0.001 |
| WHR, cm/m | 86.5 (6.6) | 93.7 (6.5) † | 83.5 (6.1) | 89.0 (6.1) * | <0.001 |
| WHtR, cm/m | 47.3 (4.9) | 54.1 (4.7) † | 48.6 (5.0) | 54.8 (5.0) * | <0.001 |
| BMI, kg/m$^2$ | 22.4 (2.8) | 26.0 (3.0) † | 22.6 (2.9) | 25.4 (3.1) * | <0.001 |
| H/W, cm/kg | 2.8 (0.4) | 2.4 (0.3) † | 2.9 (0.4) | 2.6 (0.3) * | <0.001 |
| Resting pulse rate, bpm | 79.5 (13.3) | 82.9 (13.8) † | 82.4 (12.3) | 84.4 (12.7) * | <0.001 |
| Examination results: | | | | | |
| SBP, mmHg | 135.8 (20.2) | 148.4 (18.5) † | 132.3 (20.6) | 147.6 (19.9) * | 0.028 |
| DBP, mmHg | 79.5 (12.0) | 87.5 (11.6) † | 76.6 (11.6) | 83.2 (11.5) * | <0.001 |
| FPG, mmol/L | 5.8 (1.4) | 6.7 (1.9) † | 5.7 (1.3) | 6.6 (1.9) * | 0.689 |
| P2hPG, mmol/L | 7.3 (3.3) | 9.1 (4.4) † | 7.3 (2.9) | 9.3 (4.3) * | <0.001 |
| HbA1c | 5.6 (0.8) | 6.0 (1.1) † | 5.5 (0.8) | 6.0 (1.1) * | 0.018 |
| Scr, μmmol/L | 76.6 (21.3) | 80.8 (16.5) † | 62.9 (12.2) | 65.8 (11.7) * | <0.001 |
| HDL-C, mmol/L | 1.8 (0.5) | 1.4 (0.4) † | 1.8 (0.3) | 1.5 (0.3) * | 0.017 |
| LDL-C, mmol/L | 2.9 (0.8) | 3.1 (0.8) † | 2.9 (0.8) | 3.2 (0.9) * | <0.001 |
| TC, mmol/L | 5.2 (0.9) | 5.5 (1.1) † | 5.3 (0.9) | 5.6 (1.1) * | <0.001 |
| TG, mmol/ | 1.1 (0.7) | 2.7 (2.1) † | 1.2 (0.6) | 2.2 (1.4) * | 0.039 |
| ALT, IU/L | 21.6 (13.6) | 27.8 (16.2) † | 16.4 (11.5) | 19.1 (12.8) * | <0.001 |
| AST, IU/L | 29.4 (13.5) | 32.1 (20.8) ‡ | 24.5 (10.3) | 25.1 (14.2) | <0.001 |
| GGT, IU/L | 36.5 (57.2) | 70.9 (127.4) † | 20.0 (20.8) | 27.8 (25.0) * | <0.001 |
| FIns, μU/ | 5.0 (4.7) | 8.6 (5.1) † | 6.4 (5.0) | 9.5 (5.9) * | <0.001 |
| lnHOMA-IR | 0.1 (0.6) | 0.8 (0.6) † | 0.4 (0.5) | 0.9 (0.5) * | <0.001 |
| Questionnaire results: | | | | | |
| Occupation | | | | | <0.001 |
| Government | 32 (1.4) | 24 (3.9) † | 19 (0.5) | 7 (0.4) * | |
| Medical or education | 17 (0.7) | 22 (3.6) | 29 (0.8) | 14 (0.7) | |
| Merchant | 109 (4.6) | 42 (6.8) | 120 (3.3) | 46 (2.4) | |
| Manual worker | 2013 (85.2) | 442 (71.6) | 2930 (81.6) | 1531 (78.6) | |
| Housewife | 5 (0.2) | 0 (0) | 226 (6.3) | 170 (8.7) | |
| Unemployed | 188 (8.0) | 87 (14.1) | 265 (7.4) | 180 (9.2) | |
| Culture | | | | | <0.001 |
| Illiteracy | 954 (39.7) | 201 (31.7) † | 1843 (50.8) | 1158 (59.2) * | |
| Primary or junior | 1414 (58.8) | 404 (63.7) | 1741 (48.0) | 792 (40.5) | |
| ⩾Senior high sch | 38 (1.6) | 29 (4.6) | 41 (1.1) | 5 (0.3) | |
| Smoking status | | | | | <0.001 |
| Never | 879 (35.8) | 282 (43.3) † | 3418 (95.2) | 1854 (94.3) ** | |
| Current | 777 (31.7) | 158 (24.3) | 100 (2.8) | 51 (2.6) | |
| Former | 796 (32.5) | 211 (32.4) | 71 (2.0) | 61 (3.1) | |
| Drinking status | | | | | <0.001 |
| Never | 886 (39.6) | 208 (35.1) ‡ | 2877 (89.8) | 1561 (89.0) | |
| Current | 517 (23.1) | 132 (22.3) | 138 (4.3) | 68 (3.9) | |
| Former | 832 (37.2) | 253 (42.7) | 189 (5.9) | 124 (7.1) | |
| Tea in past 1 year | | | | | 0.018 |
| Never | 1367 (58.6) | 350 (56.5) | 2895 (83.1) | 1559 (82.3) | |
| Former | 58 (2.5) | 13 (2.1) | 55 (1.6) | 35 (1.8) | |
| Current | 908 (38.9) | 257 (41.5) | 533 (15.3) | 300 (15.8) | |
| Milk-drinking habit | | | | | 0.035 |
| Yes | 1492 (68.8) | 375 (65.9) | 2260 (68.0) | 1171 (66.3) | |
| Weight change per year | | | | | 0.056 |
| Loss | 264 (11.2) | 76 (12.4) † | 399 (11.2) | 212 (11.0) | |
| Gain | 193 (8.2) | 75 (12.2) | 350 (9.8) | 219 (11.4) | |
| Stable | 1350 (57.3) | 355 (57.8) | 1987 (55.7) | 1029 (53.3) | |

(*continued on next page*)

**Table 1** (*continued*)

| | Male (n = 3177) | | Female (n = 5787) | | |
|---|---|---|---|---|---|
| | Normal (n = 2505) | MetS (n = 672) | Normal (n = 3745) | MetS (n = 2042) | *P* value for sex difference |
| Demographics: | | | | | |
| Unclear | 551 (23.4) | 108 (17.6) | 829 (23.3) | 469 (24.3) | |
| Snore | | | | | <0.001 |
| Mostly | 152 (10.1) | 58 (14.6) ‡ | 86 (3.9) | 105 (8.9) * | |
| Occasionally | 254 (16.9) | 86 (21.7) | 347 (15.7) | 224 (19.0) | |
| Never | 707 (46.9) | 162 (40.8) | 1203 (54.3) | 532 (45.2) | |
| Unclear | 393 (26.1) | 91 (22.9) | 578 (26.1) | 316 (26.8) | |
| Physical active | 176 (7.0) | 52 (7.7) | 258 (6.9) | 113 (5.5) ** | 0.132 |
| History of hypertensionsion | 117 (4.7) | 82 (12.2) † | 169 (4.5) | 257 (12.6) * | 0.051 |
| History of hyperlipidemia | 5 (0.2) | 43 (6.4) † | 3 (0.1) | 65 (3.2) * | 0.178 |
| History of diabetes | 108 (4.3) | 100 (14.9) † | 155 (4.1) | 213 (10.4) * | <0.001 |
| Age difference_child | | | | | <0.001 |
| ≥14, <22 | 46 (4.9) | 13 (4.5) | 318 (20.4) | 185 (22.5) | |
| ≥22, <30 | 794 (85.0) | 249 (87.1) | 1162 (74.4) | 598 (72.6) | |
| ≥30, ≤34 | 94 (10.1) | 24 (8.4) | 82 (5.2) | 41 (5.0) | |
| Age difference_mate | | | | | 0.048 |
| <3 | 1226 (59.9) | 328 (59.4) | 1760 (58.3) | 903 (56.0) | |
| ≥3, <10 | 784 (38.3) | 210 (38.0) | 1185 (39.3) | 660 (40.9) | |
| ≥10, ≤35 | 37 (1.8) | 14 (2.5) | 74 (2.5) | 50 (3.1) | |

Abbreviations: WC, waist circumferenc; HC, hip circumference; WHR, waist-to-hip circumferenc ratio; WHtR, waist-to-height circumference ratio; H/W, height-to-weight ratio; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; FIns, fasting plasm insulin; FPG, fasting plasma glucose;, 2h plasma glucose after 75g oral glucose tolerance test; HbA1c, Hemoglobin A1c; TG, triglycerides; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; AST, aspartate aminotransferase; ALT, alanine aminotransferase; GGT, gamma-glutamyl transferase; serum creatinine; lnHOMA-IR, natural logarithm of homeostasis model assessment of insulin resistance. Age difference_mate/child, age differences between subjects and their mate/first kid.
Data: mean (SD) for continuous variables or n (%) for categorical variables.
†$p < 0.001$ and ‡$p < 0.05$ compared with normal subjects in men, *$p < 0.001$ and **$p < 0.05$ compared with normal subjects in women.
$p$ values are calculated for difference by t-test or Mann–Whitney test and $\chi2$ test or Fisher exact test.

better in LR model 1 than the other two (model 2 and model 3), which had similar AUC values.

We analyzed the correlations of variables to screen them to develop logistic regression models and find the independent risk factors. The results showed that the correlation was high (rho = 0.852) between total cholesterol (TC) and LDL-C. Moreover, hip circumference (HP), WC, WHtR, BMI and weight revealed strong positive correlations (rho≥0.7) with each other. Correlations was also high (rho = 0.798) between lnHOMA-IR and fasting insulin. The interplay of FPG, P2hPG and HbA1c were observed in the analysis (Supplementary Table 5). After excluding those strongly correlated variables, the final model developed by step-wise logistic regression included the following factors: age, sex, drinking status, history of hyperlipidemia, RPR, SBP, HDL-C, TG, GGT, WC, height, FIns and FPG (Table 4).

Notably, in this study, we also found that former drinking habit was an important predictive factor for MetS in sex- and age-adjusted analysis (OR, 1.17; 95% CI, 1.01–1.34; Table 4) and multivariate analysis (OR, 1.35; 95% CI, 1.06–1.73). In addition, results revealed that sex was closely associated with MetS, and a higher risk of MetS was observed in female (OR, 8.84; 95% CI, 6.70–11.66). Additionally, lnHOMA-IR was a remarkable predictive factor on logistic regression analysis (OR, 2.47; 95% CI, 2.07–2.94; P < 0.0001) after adjustment for age, sex, drinking status, history of hyperlipidemia, RPR, SBP, HDL-C, TG, GGT, WC, height and FPG. It is also manifested that RPR was an important risk factor for MetS (OR, 1.01; 95% CI, 1.00–1.02; P = 0.0025) and there was a 1.0% rise in the probability of MetS with every 1-bpm increase of RPR. We also demonstrated that there was a 1.4% higher probability of MetS for each 1-cm increase of height (OR, 1.01; 95% CI, 1.00–1.03; P = 0.0536) in the multivariable logistic regression model, though the effect was marginally significant. Noteworthy, all variables of MetS components (SBP, TG, HDL, WC, FPG) were important risk factors for MetS. Moreover, it was demonstrated that the probability for MetS decrease with every 1-mmol/L increase of HDL-C (OR, 0.29; 95% CI, 0.23–0.39; Table 4).

As for sex difference, pre- and postmenopausal women both had higher risk for MetS than men after full adjustment (pre: OR, 3.77; 95% CI, 2.24–6.35; P < 0.0001; post: OR, 6.71; 95% CI, 4.48–10.05; P < 0.0001). No significant association was observed between height, RPR and MetS in total population after full adjustment. However, in the subpopulation over 58 years, there was a 3.8% higher risk of MetS for each 1-cm increase of height (OR, 1.04; 95% CI, 1.00–1.07; P = 0.0319) after full adjustment, and the effect was especially significant in men. Meanwhile, people less than 62 years had a fully adjusted 1.0% higher risk of MetS per unit increase of RPR (OR, 1.01; 95% CI, 1.00–1.02; P = 0.0358; Supplementary Table 6).

### 3.5. Comparison among LGBM and LR models

We imputed the same variables filtered out by logistic regression mentioned above into the corresponding machine learning models to compare the predictive efficacy of these two approaches. The AUC values were 0.993 for the scenario in hospital with all variables (model 1), 0.885 for the scenario in hospital with the variables of MetS components removed (model 2), 0.859 for the scenario at home without biochemical indexes (model 3) respectively by machine learning and 0.938, 0.839, 0.820 individually by logistic regression. The results revealed that accuracy, specificity in machine learning models were also higher than their corresponding LR models (Table 3).

## 4. Discussion

In this study, we developed prediction models for MetS in general population using machine learning. The model for the scenario in hospital with variables of laboratory tests, lifestyles and anthropometric measurements (Model 1) showed a best predicting efficacy and accuracy. Surprisingly, the model for the scenario in hospital with the variables of MetS components removed (Model 2) still exhibited a great performance

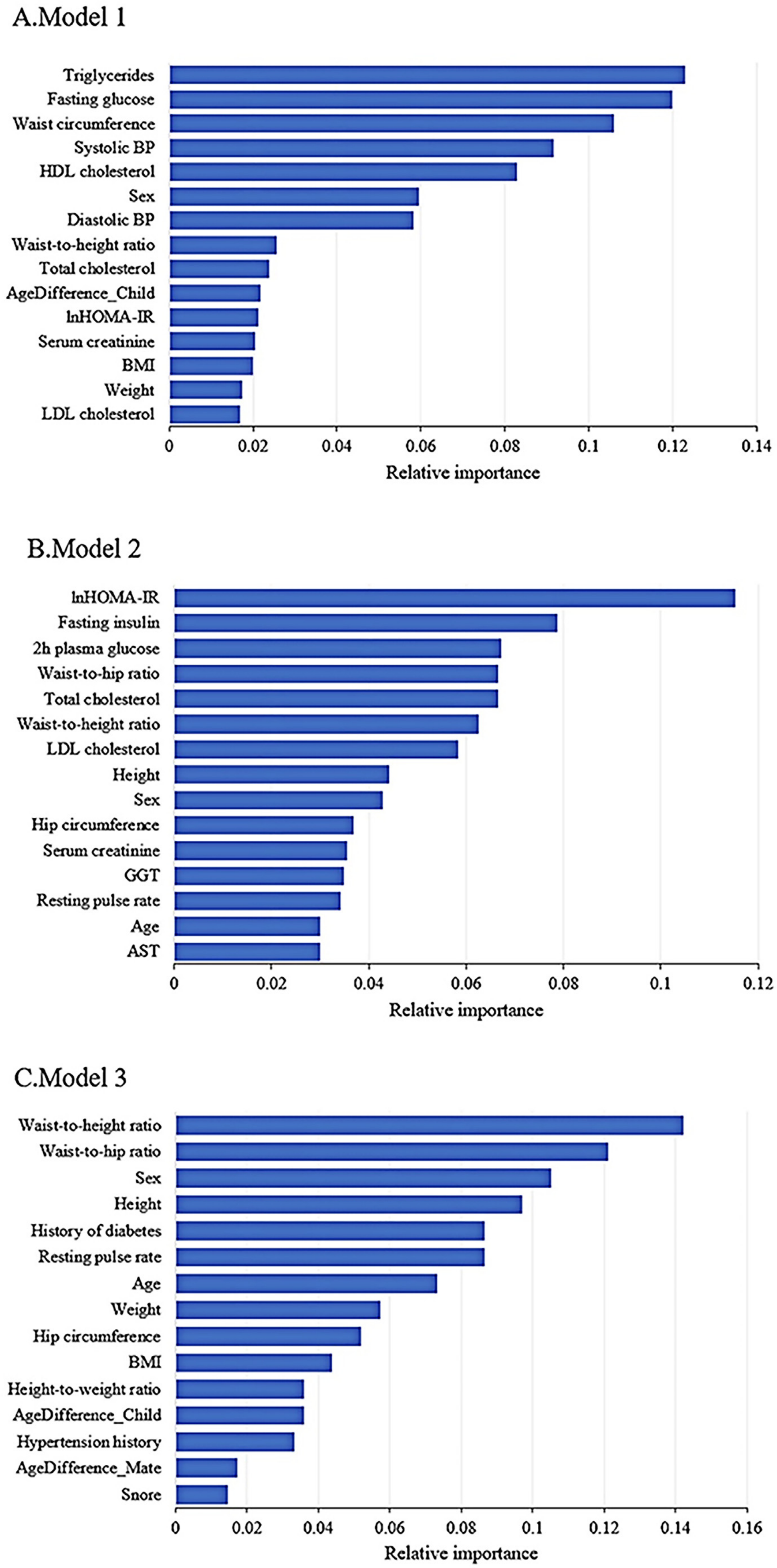

**Figure 1.** Relative importance of predictive factors according to the machine learning (LGBM) models. A. Model 1 with all 38 variables. B. Model 2 without variables which were components for the diagnosis of MetS (triglyceride, fasting blood glucose, waist circumference, systolic blood pressure, diastolic blood pressure, HDL-cholesterol) from model 1. C. Model 3 without blood biochemical indexes such as 2h-plasma glucose, lnHOMA-IR, fasting insulin, total cholesterol, LDL-cholesterol, serum creatinine, glutamate transaminase, aspartate transaminase, gamma-glutamyl transpeptidase. Relative importance represents the strength of variables' prediction ability in each model. Abbreviation: aspartate transaminase (AST), gamma-glutamyl transpeptidase (GGT).

**Table 2.** Performance comparison among various machine learning models for predicting the probability of suffering from metabolic syndrome.

| | Total | | | | Male | | | | Female | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | SP | SN | AC | AUC | SP | SN | AC | AUC | SP | SN | AC |
| Model 1 | 0.993 | 0.988 | 0.974 | 0.984 | 0.984 | 0.941 | 0.949 | 0.943 | 0.993 | 0.992 | 0.975 | 0.986 |
| Model 2 | 0.885 | 0.826 | 0.791 | 0.816 | 0.877 | 0.836 | 0.772 | 0.824 | 0.859 | 0.802 | 0.746 | 0.782 |
| Model 3 | 0.859 | 0.822 | 0.776 | 0.809 | 0.856 | 0.838 | 0.715 | 0.816 | 0.843 | 0.787 | 0.744 | 0.772 |

AUC = area under the curve.AC, accuracy; SP, specificity; SN, sensitivity. The threshold for the calculation of sensitivity and specificity was 0.5.
Attributes for each model.
Model 1: all variables; Model 2: variables in Model 1 without WC, SBP, DBP, TG, HDL, FPG; Model 3: variables in Model 2 without P2hPG, lnHOMA-IR, FIns, TC, LDL-C, Scr, AST, ALT, GGT.

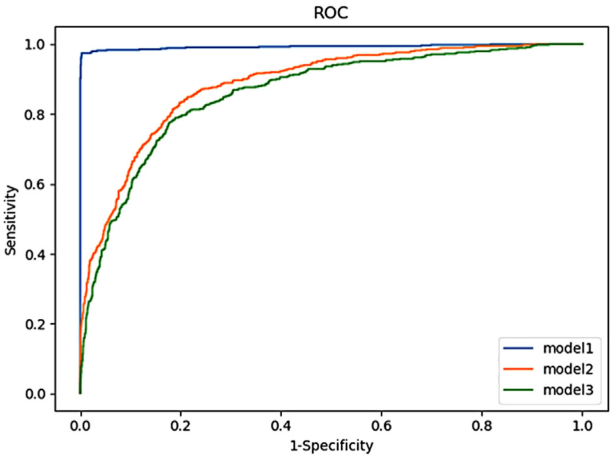

**Figure 2.** Receiver operating characteristic (ROC) curves for three models developed by machine learning in total population. Blue line represents model 1. Orange line represents model 2. Green line refers to model 3. The area under the curve represents the predictive ability of the models.

value and the Model designed for the scenario of daily life at home without biochemical indexes (Model 3) demonstrated approximately the same predicting efficacy and accuracy as model 2. Furthermore, these LGBM machine learning models had superior efficacies and accuracies than those established by logistic regression. Our findings also illustrated first time that gender, height in the subpopulation over 58 years and RPR in the subpopulation younger than 62 years were important predictors for MetS.

In recent years, the demand for more efficient healthcare delivery and health management of non-communicable chronic diseases such as MetS has been increasing in many countries including China [34]. Thus, it is very important for us to evaluate the predictors and make opportune diagnosis and intervention of these disease. Actually, logistic regression has been widely used in data analysis in this field in recent decades. However, logistic regression is being challenged with data explosion and the increasing number of features, and machine learning method has shown superior performance and accuracy in previous study ([35, 36]). Exhilaratingly, great progress has been achieved in machine learning algorithms over the past decade and these methods has been increasingly employed in many intelligent scenarios, especially in data mining for big data [37]. The combination of big data resources and new machine-learning technologies is potentially helpful to confront the difficulties by exploring new predictive strategies to detect potential patients such as MetS timely for proper diagnosis and intervention [38]. The prediction models developed by machine learning are able to take lots of information into account to reach a decision, which is very similar to the conventional way adopted by physicians to make diagnostic decisions [39]. Inspiringly, a prediction model using artificial neural network (ANN), which is a common machine learning algorithm based on the structure of the brain tissue [40, 41], is developed by Darko Ivanovic et al. [42] for the diagnosis of MetS, implying that the prediction of MetS diagnosis using machine learning seems feasible and is likely to be implemented in clinical practice. Later, prediction models using machine learning was developed to identify the risk of suffering from MetS in Japan for individuals classified as high-risk who are not able to participate in a health intervention program [35] and in Korea for nonobese population [36]. However, there is no developed model by machine learning to predict the probability of suffering from MetS based on large population study. Remarkably, in statistics and machine learning, ensemble methods were developed using multiple learning algorithms to obtain better predictive performance than that could be obtained from any of the constituent learning algorithms alone [43]. Among them, the method of decision tree (DT) has an inherent function that is ranking the importance of variables [44], which are of great help for us the develop the prediction model by assessing the priority of features in the model. Hyper parameter optimization was applied to LGBM and XGBoost algorithms to increase classification performance. However, LGBM classifier surpassed XGBoost classifier in terms of performance and processing time [45]. Therefore, the present study used LGBM method, which is enhanced framework of DT ensemble learning and shows strong predictive efficiency and robustness. Herein, considering that the prevention screening tool is expected to be universal for general population and the biomedical indexes in these models are a little bit inaccessible for self-assessment in daily life, in the present study, we constructed prediction models with easily-obtained indicators at home for self-assessment with a large, random sampling, population-based study of 8964 individuals in the urban and rural areas of central China, and potential risk factors are also involved as comprehensive as possible for further modelling and analysis.

In the present study, the reliability and accuracy of the predictive models for MetS was confirmed by LGBM model 1 with all MetS components used to diagnose MetS, in which the performance was almost

**Table 3.** Comparison among LGBM and logistic regression methods.

| | Logistic regression | | | | LGBM | | | |
|---|---|---|---|---|---|---|---|---|
| | AC | AUC | SN | SP | AC | AUC | SN | SP |
| Model 1 | 0.857 | 0.938 | 0.907 | 0.835 | 0.985 | 0.993 | 0.975 | 0.989 |
| Model 2 | 0.745 | 0.839 | 0.822 | 0.712 | 0.816 | 0.885 | 0.791 | 0.826 |
| Model 3 | 0.746 | 0.820 | 0.755 | 0.742 | 0.810 | 0.860 | 0.777 | 0.822 |

Variables included in three LGBM models developed above.

**Table 4.** Association between variables and MetS by unadjusted analysis, age- and sex-adjusted analysis, and multivariate stepwise logistic regression analysis.

| Baseline variable | Unadjusted Covariates | Age- and Sex-Adjusted Covariates | Fully Adjusted Model (stepwise regression) |
|---|---|---|---|
| Age, years | 1.02 (1.02, 1.03) * | | 1.02 (1.01,1.03) * |
| Women | 2.03 (1.84, 2.25) * | | 8.84 (6.70,11.66) * |
| Smoking status | | | |
| Never | Reference | Reference | |
| Current | 0.50 (0.43, 0.58) * | 0.74 (0.62, 0.88) * | |
| Former | 0.64 (0.55, 0.74) * | 0.97 (0.82, 1.16) 0.7619 | |
| Drinking status | | | |
| Never | Reference | Reference | Reference |
| Current | 0.66 (0.56, 0.77) * | 0.95 (0.80, 1.13) 0.5710 | 0.79 (0.58,1.08)0.1406 |
| Former | 0.78 (0.69, 0.88) * | 1.17 (1.01, 1.34) 0.0311 | 1.35 (1.06,1.73)0.0150 |
| Milk-drinking habit (no) | 1.11 (1.01, 1.22) 0.0347 | 1.10 (1.00, 1.21) 0.0506 | |
| Weight change per year | | | |
| Loss | Reference | Reference | |
| Gain | 1.17 (0.97, 1.41) * | 1.21 (1.00, 1.47) 0.0491 | |
| Stable | 0.91 (0.79, 1.05) * | 0.91 (0.79, 1.06) 0.2211 | |
| Unclear | 0.91 (0.77, 1.07) * | 0.86 (0.73, 1.01) 0.0647 | |
| Snore | | | |
| Mostly | Reference | Reference | |
| Occasionally | 0.80 (0.67, 0.95) 0.0116 | 0.74 (0.62, 0.88) * | |
| Never | 0.63 (0.54, 0.74) * | 0.56 (0.48, 0.65) * | |
| Unclear | 0.66 (0.56, 0.78) * | 0.59 (0.50, 0.70) * | |
| Resting pulse rate, bpm | 1.02 (1.01, 1.02) * | 1.01 (1.01, 1.02) * | 1.01 (1.00,1.02)0.0025 |
| SBP, mmHg | 1.03 (1.03, 1.04) * | 1.03 (1.03, 1.04) * | 1.04 (1.03,1.04) * |
| DBP, mmHg | 1.05 (1.04, 1.05) * | 1.05 (1.05, 1.06) * | |
| FPG, mmol/L | 1.49 (1.43, 1.56) * | 1.48 (1.42, 1.55) * | 1.29 (1.23,1.36) * |
| P2hPG, mmol/L | 1.16 (1.14, 1.17) * | 1.15 (1.14, 1.17) * | |
| HbA1c, % | 1.66 (1.56, 1.76) * | 1.64 (1.54, 1.74) * | |
| Scr, μmol/L | 1.00 (1.00, 1.01) 0.0043 | 1.01 (1.01, 1.02) * | |
| HDL-C, mmol/L | 0.10 (0.08, 0.11) * | 0.07 (0.06, 0.08) * | 0.29 (0.23,0.39) * |
| LDL-C, mmol/L | 1.47 (1.39, 1.55) * | 1.42 (1.34, 1.50) * | |
| TC, mmol/L | 1.39 (1.33, 1.46) * | 1.34 (1.28, 1.40) * | |
| TG, mmol/L | 5.13 (4.71, 5.58) * | 5.20 (4.77, 5.67) * | 3.87 (3.42,4.38) * |
| ALT, IU/L | 1.02 (1.01, 1.02) * | 1.03 (1.02, 1.03) * | |
| AST, IU/L | 1.00 (1.00, 1.01) 0.2574 | 1.01 (1.00, 1.01) 0.0015 | |
| GGT, IU/L | 1.01 (1.00, 1.01) * | 1.01 (1.01, 1.01) * | 1.00 (1.00,1.00)0.0120 |
| FIns, μU/ml | 1.24 (1.22, 1.26) * | 1.23 (1.22, 1.25) * | 1.01 (1.00,1.03)0.0743 |
| lnHOMA-IR | 7.89 (7.08, 8.80) * | 7.82 (7.00, 8.73) * | |
| Weight, kg | 1.09 (1.08, 1.09) * | 1.14 (1.13, 1.14) * | |
| Height, cm | 0.99 (0.98, 0.99) * | 1.04 (1.03, 1.05) * | 1.01 (1.00, 1.03) 0.0536 |
| WC, cm | 1.17 (1.16, 1.17) * | 1.19 (1.18, 1.20) * | 1.17 (1.16,1.19) * |
| HC, cm | 1.15 (1.14, 1.16) * | 1.15 (1.14, 1.16) * | |
| WHR, cm/m | 1.30 (1.28, 1.31) * | 1.29 (1.28, 1.31) * | |
| WHtR, cm/m | 1.14 (1.13, 1.15) * | 1.17 (1.16, 1.18) * | |
| BMI, kg/m$^2$ | 1.40 (1.37, 1.43) * | 1.42 (1.40, 1.45) * | |
| H/W, cm/kg | 0.07 (0.06, 0.08) * | 0.04 (0.03, 0.05) * | |
| History of diabetes | 3.14 (2.81, 3.51) * | 3.15 (2.81, 3.53) * | |
| History of hypertension | 2.98 (2.53, 3.51) * | 2.77 (2.34, 3.27) * | |
| History of hyperlipidemia (no) | 0.03 (0.02, 0.06) * | 0.03 (0.01, 0.05) * | 0.01 (0.00,0.03) * |

Data are presented as the OR (95% CI) $p$ value. *$p < 0.001$. The analysis initially considered all the covariates that were statistically significant on age- and sex-adjusted analysis.

perfect. The existence of TG, FPG, WC, SBP, HDL-C and DBP, which all present high relative importance in this model, makes LGBM model 1 acquire excellent prediction accuracy. Notably, a relatively good efficacy was observed in LGBM model 2 without six indexes of MetS components, which is probably due to the great contributions of other crucial risk factors associated with MetS. It suggests that model 2 is feasible in the hospital to predict MetS without diagnostic variables of MetS. In this model, it is demonstrated that HOMA-IR and FIns are efficacious factors in predicting MetS as well as obesity, dyslipidemia and hyperglycemia. These results implied that insulin resistance is pivotal and dominant in the association with MetS. That insulin resistance mediates the increases in gluconeogenesis and lipogenesis [46], and consequently induces hyperglycemia and dyslipidemia, might be important mechanistic explanations. Herein, early evaluation of insulin sensitivity or fasting plasma insulin level may be of great help for clinicians to predict the probability of suffering from MetS and identify undiagnosed MetS patients timely from the high-probability individuals, though the complexity might be far beyond our current understanding and further research is necessary.

Most physicians are convinced that laboratory test results are indispensable and the other indicators not generated from clinical laboratories, except WC and SBP/DBP, are of limited value in prediction or diagnosis of MetS [39, 47]. Marvelously, LGBM model 3 developed in the present study, trained without any biomedical indexes and using non-invasive variables (routine lifestyle indicators that are the most frequently used by people), was almost the same effective in predicting MetS as LGBM model 2. It implied that the LGBM model 3 can help clinicians to screen MetS patient smore easily, not always relying on the laboratory assays, especially in primary care clinics or hospitals, in which clinical laboratory testing is not available. Moreover, the LGBM model 3 could be particularly useful for individuals who are not medical professionals, to evaluate the possibility of suffering from MetS in their daily life for opportune further medical consult and decision making [39].

Our result illustrated that these models developed by LGBM exhibited higher AUC value, sensitivity, specificity and accuracy, indicating a greater predicting efficacy compared with the models generated by logistic regression method, which is consistent with the work of Akihiro Shimoda et al. [35]. The superiorities of the LGBM models are faster training efficiency, higher accuracy and it is considered that machine learning models has an inherent excellence to assess complex interaction effects between all features [21], compared to the models developed by logistic regression. Based on this fact, Yang et al. concerns with a machine learning-aided longitudinal study on risk prediction of MetS by using a total of three consecutive years examination records of 67,730 individuals which is shown that the proposed risk prediction model yields a higher performance in comparison with the state-of-the-art methods [48]. These results suggested that a machine learning approach could work very efficiently with easily-obtained information imputed, which is consistent with the results of Darko Ivanovic et al. [42]. Furthermore, it might be greatly improved to become an integral part of medical professional systems [39] and provide new ideas for developing healthcare decision support systems based on large-scale, digital databases of patient information that might change the model of healthcare.

Moreover, it is demonstrated that some potential variables might have implications for MetS in our LGBM models. Height was shown to be a risk factor in elderly people in the present study. It is reported by Heymsfield et al. [49] that body fat percentage is scaled positively to height in Mexican American men in their study and tall Mexican American men in National Health and Nutrition Examination Survey (NHANES) have a higher fat percentage compared with their short counterparts, implying that fat would associate with height. Arvedsen et al. studied that 24-h ambulatory mean arterial pressure significantly increased in taller males [50]. Furthermore, it is recently reported that body height is closely related with the risk of cardiovascular diseases [51], atrial fibrillation and mitral valve prolapse [52]. These studies suggested that height might be an unexplored contributing factor to metabolic abnormalities which guide the search for much-needed effective therapies in this population. Our results also indicated that RPR is likely to be a new predictor for MetS. In general, resting heart rate and resting pulse rate are consistent in people without severe heart diseases. It is reported that higher RHR is associated with increases in fat, blood pressure and serum glucose, afferent neuronal signals to the brain leading to modulation of sympathetic tonus [53, 54], which is attributed to the elevation of level of oxidative stress and pro-inflammation cytokines, decreased arterial compliance and distensibility, and consequently results in metabolic and cardiovascular disorders [55]. It is reported that elevated RHR is associated with increased systemic inflammation and endothelial dysfunction [56], which are the major features of the metabolic and cardiovascular abnormalities including MetS and cardiometabolic disorders [57]. A meta-analysis reported that the risk of suffering from MetS may elevated with the increases of resting heart rate (RHR) [58], which is concordant with our findings in this study. Anyhow, the exact mechanisms were unclear and further research is needed.

Noteworthy, it is demonstrated that WHtR and WHR, which were easy-to-obtain and cost-effective anthropometric markers of obesity, were high in the ranking of variable weights in model 3. It is reported that WHtR was a good screening tool for MetS [59, 60] and WHR was the anthropometric index that showed the highest predictive value for MetS components [61], which might be important explanations of our findings in this study. Additionally, it is also illustrated in this study by using LGBM that sex difference is potentially predictive of the probability of suffering from MetS. Actually, the gender differences in the prevalence of metabolic syndrome vary in different studies. Our regression analysis in the present study showed that being female was a risk factor for MetS, which is consistent with the results of some previous studies [4, 9, 11, 62, 63]. More and more studies hypothesize that menopause-related hormonal changes affect the prevalence of MetS [64, 65], in which the physiological changes during menopause leading to hormonal homeostatic dysregulation [66] may explain that the female has higher risk of MetS. Nevertheless, our study shows that premenopausal women also have higher risk of suffering from MetS than men, which suggests that hormonal changes in the menopause may not be the only key risk factor in the sex-specific differences in the prevalence of MetS. Differences in regional and ethnic variations, socio-economic, cultural behaviors, education levels, dietary habits disparities and physical activity levels are likely to influence this gender differences in MetS prevalence [4, 62, 63, 67]. It is reported in NHANES that a greater relative increase in the prevalence of MetS have was observed in women compared to men (22.8 vs. 11.2%) in the USA [67]. Low education, low socio-economic status and physical inactive significantly increased the odds for MetS [64, 68] while women were prone to these traits in their study. The gender-specific prevalence of MetS might also be attributed to other important factors such as genetic differences [69]. Anyhow, gender differences in the prevalence of MetS should be taken into account in clinical practice or research, which would be helpful in the early detection of MetS in the general Chinese population aged 40–75 years.

Our results showed components of MetS had a high rate of missed diagnoses previously, and the prediction models developed by machine learning could be helpful for the screening of these undetected disorders and have high efficacy and accuracy in predicting the probability of suffering from MetS. These approaches are probably helpful for individuals, who are unable to participate in a health check-up promptly, to have a preliminary knowledge of their probabilities of suffering from MetS and get opportune healthcare such as well-timed diagnosis and treatment if needed. These models are potentially widely used for convenient and non-invasive self-assessment in general population in their daily life without any medical tests, which could be of great help in the prevention and control of MetS for its remarkable improvement of accessibility in health estimation. Furthermore, this model could help individuals at high-risk of MetS pay more attention and further initiate health examinations and intervention with healthy lifestyle.

### 4.1. Strengths and limitations

In this study, we constructed different models using machine learning technique in the set of various scenarios which is probably of great help to predict the probability of suffering from MetS in population and to fulfil early diagnosis and timely intervention, which is very important for their healthcare and health management. Additionally, more variables were included in the analysis and more predictors could be filtered out by LGBM compared with that by logistic regression. Moreover, our results demonstrated that the LGBM machine learning technique is of great help to improve the efficacy and accuracy in the predicting the probability of suffering from MetS in individuals.

Yet, it should be noted that the present study did not include people who were under 40 or more than 75 years old, in whom our findings herein were probably not applicable. Actually, there were still some missing data in the large population-based study which might cause some bias, though we had tried our best to collect as much information as we can. However, the number of the cases are 8964 individuals from a large, random sampling, population-based study in the urban and rural areas which is helpful to make our models more stable and credible. Additionally, since machine learning methods have been developing rapidly, it is likely that the models instigated in the present study could be further improved and more accurate and efficacious models could be developed to predict MetS in the future.

Notwithstanding these limitations, the prediction models developed by machine learning in the present study were effective in evaluating the probability of suffering from MetS, and presented prominent predicting efficacy and accuracy for distinguishing potential patients of MetS. These models would be of great help to predict the probability of suffering from MetS in individuals accurately, as well as subsequently fulfil well-timed diagnosis and initiate intervention as early as possible. Moreover, our results indicated that gender, height and RPR were important factors in predicting the probability of suffering from MetS.

## Declarations

### *Author contribution statement*

1 - Conceived and designed the experiments
2 - Performed the experiments
3 - Analyzed and interpreted the data
4 - Contributed reagents, materials, analysis tools or data
5 - Wrote the paper.

### *Funding statement*

Dr. Lulu Chen was supported by National Natural Science Foundation of China [81471069].

### *Data availability statement*

The authors do not have permission to share data.

### *Declaration of interest's statement*

The authors declare no conflict of interest.

### *Additional information*

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e12343.

## Acknowledgements

The authors thank all research staff from the Department of Endocrinology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, the local health systems for their collection of data and blood samples, and all the staff of the REACTION Study Group. We also thank all the study participants for their participation and contribution to this study.

## References

[1] H.M. Lakka, D.E. Laaksonen, T.A. Lakka, L.K. Niskanen, E. Kumpusalo, J. Tuomilehto, J.T. Salonen, The metabolic syndrome and total and cardiovascular disease mortality in middle-aged men, JAMA 288 (2002) 2709–2716.

[2] S.M. Grundy, J.I. Cleeman, S.R. Daniels, K.A. Donato, R.H. Eckel, B.A. Franklin, D.J. Gordon, R.M. Krauss, P.J. Savage, S.C. Smith Jr., J.A. Spertus, F. Costa, American heart A, national heart L, blood I. Diagnosis and management of the metabolic syndrome: an American heart association/national heart, lung, and blood institute scientific statement, Circulation 112 (2005) 2735–2752.

[3] H. Beltran-Sanchez, M.O. Harhay, M.M. Harhay, S. McElligott, Prevalence and trends of metabolic syndrome in the adult U.S. population, 1999-2010, J. Am. Coll. Cardiol. 62 (2013) 697–703.

[4] B. Xi, D. He, Y.H. Hu, D.H. Zhou, Prevalence of metabolic syndrome and its influencing factors among the Chinese adults: the China Health and Nutrition Survey in 2009, Prev. Med. 57 (2013) 867–871.

[5] J.K.K. Vishram, A. Borglykke, A.H. Andreasen, J. Jeppesen, H. Ibsen, T. Jorgensen, L. Palmieri, S. Giampaoli, C. Donfrancesco, F. Kee, G. Mancia, G. Cesana, K. Kuulasmaa, V. Salomaa, S. Sans, J. Ferrieres, J. Dallongeville, S. Soderberg, D. Arveiler, A. Wagner, H. Tunstall-Pedoe, W. Drygas, M.H. Olsen, M. Project, Impact of age and gender on the prevalence and prognostic importance of the metabolic syndrome and its components in Europeans. The MORGAM prospective cohort project, PLoS One 9 (2014).

[6] A. Engin, The definition and prevalence of obesity and metabolic syndrome, Adv. Exp. Med. Biol. 960 (2017) 1–17.

[7] K.G. Lim, W.K. Cheah, A review of metabolic syndrome research in Malaysia, Med. J. Malaysia 71 (2016) 20–28.

[8] S.M. Grundy, Metabolic syndrome: a multiplex cardiovascular risk factor, J. Clin. Endocrinol. Metab. 92 (2007) 399–404.

[9] W.Z. Li, F.J. Song, X.J. Wang, L.D. Wang, D.M. Wang, X.X. Yin, S.Y. Cao, Y.H. Gong, W. Yue, F. Yan, H. Zhang, Z.J. Sheng, Z.H. Wang, Z.X. Lu, Prevalence of metabolic syndrome among middle-aged and elderly adults in China: current status and temporal trends, Ann. Med. 50 (2018) 345–353.

[10] G. Pucci, R. Alcidi, L. Tap, F. Battista, F. Mattace-Raso, G. Schillaci, Sex- and gender-related prevalence, cardiovascular risk and therapeutic approach in metabolic syndrome: a review of the literature, Pharmacol. Res. 120 (2017) 34–42.

[11] R. Li, W.C. Li, Z.J. Lun, H.P. Zhang, Z. Sun, J.S. Kanu, S. Qiu, Y. Cheng, Y.W. Liu, Prevalence of metabolic syndrome in mainland China: a meta-analysis of published studies, BMC Publ. Health 16 (2016).

[12] F. Karimi-Alavijeh, S. Jalili, M. Sadeghi, Predicting metabolic syndrome using decision tree and support vector machine methods, Arya Atherosclerosis 12 (2016) 146–152.

[13] N. Santoro, A. Amato, A. Grandone, C. Brienza, P. Savarese, N. Tartaglione, P. Marzuillo, L. Perrone, E. Miraglia Del Giudice, Predicting metabolic syndrome in obese children and adolescents: look, measure and ask, Obesity facts 6 (2013) 48–56.

[14] R.K. Simmons, A.H. Harding, N.J. Wareham, S.J. Griffin, Do simple questions about diet and physical activity help to identify those at risk of Type 2 diabetes? Diabet. Med. J. British Diabetic Association 24 (2007) 830–835.

[15] I. Contreras, J. Vehi, Artificial intelligence for diabetes management and decision support: literature review, J. Med. Internet Res. 20 (2018), e10775.

[16] U. Schmidt-Erfurth, A. Sadeghipour, B.S. Gerendas, S.M. Waldstein, H. Bogunovic, Artificial intelligence in retina, Prog. Retin. Eye Res. 67 (2018) 1–29.

[17] B.J. Erickson, P. Korfiatis, Z. Akkus, T.L. Kline, Machine learning for medical imaging(1), Radiographics 37 (2017) 505–515.

[18] D.S.W. Ting, C.Y.L. Cheung, G. Lim, G.S.W. Tan, N.D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I.Y.S. Yeo, S.Y. Lee, E.Y.M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N.C. Tan, E.A. Finkelstein, E.L. Lamoureux, I.Y. Wong, N.M. Bressler, S. Sivaprasad, R. Varma, J.B. Jonas, M.G. He, C.Y. Cheng, G.C.M. Cheung, T. Aung, W. Hsu, M.L. Lee, T.Y. Wong, Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes, JAMA, J. Am. Med. Assoc. 318 (2017) 2211–2223.

[19] G. Heller, A measure of explained risk in the proportional hazards model, Biostatistics 13 (2012) 315–325.

[20] J.F. Ludvigsson, E. Andersson, A. Ekbom, M. Feychting, J.L. Kim, C. Reuterwall, M. Heurgren, P.O. Olausson, External review and validation of the Swedish national inpatient register, BMC Publ. Health 11 (2011).

[21] G.L. Ke, Q. Meng, T. Finley, T.F. Wang, W. Chen, W.D. Ma, Q.W. Ye, T.Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, Adv Neur In (2017) 30.

[22] Y. Bi, J. Lu, W. Wang, Y. Mu, J. Zhao, C. Liu, L. Chen, L. Shi, Q. Li, Q. Wan, S. Wu, T. Yang, L. Yan, Y. Liu, G. Wang, Z. Luo, X. Tang, G. Chen, Y. Huo, Z. Gao, Q. Su, Z. Ye, Y. Wang, G. Qin, H. Deng, X. Yu, F. Shen, L. Chen, L. Zhao, J. Zhang, J. Sun, M. Dai, M. Xu, Y. Xu, Y. Chen, S. Lai, Z.T. Bloomgarden, D. Li, G. Ning, Cohort profile: risk evaluation of cancers in Chinese diabetic individuals: a longitudinal (REACTION) study, J. Diabetes 6 (2014) 147–157.

[23] R. Li, W. Li, Z. Lun, H. Zhang, Z. Sun, J.S. Kanu, S. Qiu, Y. Cheng, Y. Liu, Prevalence of metabolic syndrome in Mainland China: a meta-analysis of published studies, BMC Publ. Health 16 (2016) 296.

[24] W. Tao, Z. Zeng, H. Dang, P. Li, L. Chuong, D. Yue, J. Wen, R. Zhao, W. Li, G. Kominski, Towards universal health coverage: achievements and challenges of 10 years of healthcare reform in China, BMJ global health 5 (2020), e002087.

[25] N. Zhang, X. Hu, Q. Zhang, P. Bai, M. Cai, T.S. Zeng, J.Y. Zhang, S.H. Tian, J. Min, H.T. Huang, J. Zheng, M.M. Peng, M.J. Li, L.L. Chen, Non-high-density lipoprotein cholesterol:High-density lipoprotein cholesterol ratio is an independent risk factor for diabetes mellitus: results from a population-based cohort study, J. Diabetes 10 (2018) 708–714.

[26] X. Hu, Q. Zhang, M. Zhang, X. Yang, T.S. Zeng, J.Y. Zhang, J. Zheng, W. Kong, J. Min, S.H. Tian, R. Zhu, Z. Yuan, C. Wu, L.L. Chen, Tannerella forsythia and coating color on the tongue dorsum, and fatty food liking associate with fat accumulation and insulin resistance in adult catch-up fat, Int. J. Obes. 42 (2018) 121–128.

[27] K.G. Alberti, R.H. Eckel, S.M. Grundy, P.Z. Zimmet, J.I. Cleeman, K.A. Donato, J.C. Fruchart, W.P. James, C.M. Loria, S.C. Smith Jr., Harmonizing the metabolic syndrome: a joint interim statement of the international diabetes federation task force on epidemiology and prevention; national heart, lung, and blood institute; American heart association; world heart federation; international atherosclerosis society; and international association for the study of obesity, Circulation 120 (2009) 1640–1645.

[28] B. Bozkurt, D. Aguilar, A. Deswal, S.B. Dunbar, G.S. Francis, T. Horwich, M. Jessup, M. Kosiborod, A.M. Pritchett, K. Ramasubbu, C. Rosendorff, C. Yancy, Contributory risk and management of comorbidities of hypertension, obesity, diabetes mellitus, hyperlipidemia, and metabolic syndrome in chronic heart failure: a scientific statement from the American heart association, Circulation 134 (2016) e535–e578.

[29] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232.

[30] M.H. Kutner, C. Nachtsheim, J. Neter, Applied Linear Regression Models, McGraw-Hill/Irwin, Boston ; New York, 2004.

[31] F.E. Harrell Jr., K.L. Lee, D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Stat. Med. 15 (1996) 361–387.

[32] Y.S. Su, A. Gelman, J. Hill, M. Yajima, Multiple imputation with diagnostics (mi) in R: opening windows into the black box, J. Stat. Software 45 (2011) 1–31.

[33] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, Wiley-Interscience, 2004. Hoboken, N.J.

[34] Q. Yao, Y. Tian, P.F. Li, L.L. Tian, Y.M. Qian, J.S. Li, Design and development of a medical big data processing system based on Hadoop, J. Med. Syst. 39 (2015) 23.

[35] A. Shimoda, D. Ichikawa, H. Oyama, Prediction models to identify individuals at risk of metabolic syndrome who are unlikely to participate in a health intervention program, Int. J. Med. Inf. 111 (2018) 90–99.

[36] E.K. Choe, H. Rhee, S. Lee, E. Shin, S.W. Oh, J.E. Lee, S.H. Choi, Metabolic syndrome prediction using machine learning models with genetic and clinical information from a nonobese healthy population, Genomics Inform 16 (2018) e31.

[37] M.I. Jordan, T.M. Mitchell, Machine learning: trends, perspectives, and prospects, Science 349 (2015) 255–260.

[38] L. Zhang, H. Wang, Q. Li, M.H. Zhao, Q.M. Zhan, Big data and medical research in China, BMJ 360 (2018) j5910.

[39] G. Guncar, M. Kukar, M. Notar, M. Brvar, P. Cernelc, M. Notar, M. Notar, An application of machine learning to haematological diagnosis, Sci Rep-Uk 8 (2018).

[40] R.W. Filice, S.K. Frantz, Effectiveness of deep learning algorithms to determine laterality in radiographs, J. Digit. Imag. 32 (2019) 656–664.

[41] J. Jing, M. Ge, Z. Yang, P. Li, Spatial distribution characteristics of tumor marker CA724 reference values in China, Cancer Med. 8 (2019) 4465–4474.

[42] D. Ivanovic, A. Kupusinac, E. Stokic, R. Doroslovacki, D. Ivetic, ANN prediction of metabolic syndrome: a complex puzzle that will be completed, J. Med. Syst. 40 (2016).

[43] Zhang C, Ma Y. Ensemble Machine Learning || Ensemble Learning.

[44] Changwon Y, Luis R, Journal LJJIN. Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine. 18:50-

[45] I.F. Kilincer, F. Ertam, A. Sengur, A comprehensive intrusion detection framework using boosting algorithms, Comput. Electr. Eng. 100 (2022), 107869.

[46] Y. Rochlani, N.V. Pothineni, S. Kovelamudi, J.L. Mehta, Metabolic syndrome: pathophysiology, management, and modulation by natural compounds, Ther Adv Cardiovasc Dis 11 (2017) 215–225.

[47] Y. Luo, P. Szolovits, A.S. Dighe, J.M. Baron, Using machine learning to predict laboratory test results, Am. J. Clin. Pathol. 145 (2016) 778–788.

[48] H. Yang, B. Yu, P. Ou, X. Li, X. Lai, G. Zhang, H. Zhang, Machine learning-aided risk prediction for metabolic syndrome based on 3 years study, Sci. Rep. 12 (2022) 2248.

[49] S.B. Heymsfield, M.S. Heo, D. Thomas, A. Pietrobelli, Scaling of body composition to height: relevance to height-normalized indexes, Am. J. Clin. Nutr. 93 (2011) 736–740.

[50] S.K. Arvedsen, M. Damgaard, P. Norsk, Body height and blood pressure regulation in humans during anti-orthostatic tilting, Am. J. Physiol. Regul. Integr. Comp. Physiol. 302 (2012) R984–989.

[51] D. Montero, C. Diaz-Canestro, Body height is inversely associated with left ventricular end-diastolic pressure in heart failure with preserved ejection fraction, European J. Prevent. Cardiol. 27 (2020) 1116–1118.

[52] S.W. Rosenbush, J.M. Parker, Height and heart disease, Rev. Cardiovasc. Med. 15 (2014) 102–108.

[53] A. Niijima, Afferent signals from leptin sensors in the white adipose tissue of the epididymis, and their reflex effect in the rat, J. Auton. Nerv. Syst. 73 (1998) 19–25.

[54] T. Miyawaki, M. Abe, K. Yahata, N. Kajiyama, H. Katsuma, N. Saito, Contribution of visceral fat accumulation to the risk factors for atherosclerosis in non-obese Japanese, Intern. Med. (Tokyo) 43 (2004) 1138–1144.

[55] P. Palatini, Heart rate as an independent risk factor for, cardiovascular disease - current evidence and basic mechanisms, Drugs 67 (2007) 3–13.

[56] D. Nanchen, D.J. Stott, J. Gussekloo, S.P. Mooijaart, R.G. Westendorp, J.W. Jukema, P.W. Macfarlane, J. Cornuz, N. Rodondi, B.M. Buckley, I. Ford, N. Sattar, A.J. de Craen, Resting heart rate and incident heart failure and cardiovascular mortality in older adults: role of inflammation and endothelial dysfunction: the PROSPER study, Eur. J. Heart Fail. 15 (2013) 581–588.

[57] N. Bernardes, P. Ayyappan, K. De Angelis, A. Bagchi, G. Akolkar, D. da Silva Dias, A. Belló-Klein, P.K. Singal, Excessive consumption of fructose causes cardiometabolic dysfunctions through oxidative stress and inflammation, Can. J. Physiol. Pharmacol. 95 (2017) 1078–1090.

[58] X.J. Liu, X.P. Luo, Y. Liu, X.Z. Sun, C.Y. Han, L. Zhang, B.Y. Wang, Y.C. Ren, Y. Zhao, D.D. Zhang, D.S. Hu, M. Zhang, Resting heart rate and risk of metabolic syndrome in adults: a dose-response meta-analysis of observational studies, Acta Diabetol. 54 (2017) 223–235.

[59] J.Y. Kim, S. Oh, M.R. Chang, Y.G. Cho, K.H. Park, Y.J. Paek, S.H. Yoo, J.J. Cho, I.D. Caterson, H.J. Song, Comparability and utility of body composition measurement vs. anthropometric measurement for assessing obesity related health risks in Korean men, Int. J. Clin. Pract. 67 (2013) 73–80.

[60] S.X. Guo, X.H. Zhang, J.Y. Zhang, J. He, Y.Z. Yan, J.L. Ma, R.L. Ma, H. Guo, L.T. Mu, S.G. Li, Q. Niu, D.S. Rui, M. Zhang, J.M. Liu, K. Wang, S.Z. Xu, X. Gao, Y.S. Ding, Visceral adiposity and anthropometric indicators as screening tools of metabolic syndrome among low income rural adults in xinjiang, Sci Rep-Uk 6 (2016).

[61] J.S. Perona, J. Schmidt-RioValle, B. Rueda-Medina, M. Correa-Rodriguez, E. Gonzalez-Jimenez, Waist circumference shows the highest predictive value for metabolic syndrome, and waist-to-hip ratio for its components, in Spanish adolescents, Nutr. Res. 45 (2017) 38–45.

[62] S.S. Yu, X.F. Guo, H.M. Yang, L.Q. Zheng, Y.X. Sun, An update on the prevalence of metabolic syndrome and its associated factors in rural northeast China, BMC Publ. Health 14 (2014).

[63] Y.L. Zhao, H. Yan, R.H. Yang, Q. Li, S.N. Dang, Y.Y. Wang, Prevalence and determinants of metabolic syndrome among adults in a rural area of Northwest China, PLoS One 9 (2014).

[64] G. Pucci, R. Alcidi, L. Tap, F. Battista, F. Mattace-Raso, G. Schillaci, Sex- and gender-related prevalence, cardiovascular risk and therapeutic approach in metabolic syndrome: a review of the literature, Pharmacol. Res. 120 (2017) 34–42.

[65] A.D. Pradhan, Sex differences in the metabolic syndrome: implications for cardiovascular health in women, Clin. Chem. 60 (2014) 44–52.

[66] I. Janssen, L.H. Powell, S. Crawford, B. Lasley, K. Sutton-Tyrrell, Menopause and the metabolic syndrome - the study of women's health across the nation, Arch. Intern. Med. 168 (2008) 1568–1575.

[67] A. Mozumdar, G. Liguori, Persistent increase of prevalence of metabolic syndrome among US adults: NHANES III to NHANES 1999-2006, Diabetes Care 34 (2011) 216–219.

[68] J.X. Moore, N. Chaudhary, T. Akinyemiju, Metabolic syndrome prevalence by race/ethnicity and sex in the United States, national health and nutrition examination survey, 1988-2012, Prev. Chronic Dis. 14 (2017).

[69] C.X. Wang, B. Wang, H.J. He, X.T. Li, D.Y. Wei, J.H. Zhang, M.J. Ma, L. Pan, T. Yu, F. Xue, L. Li, G.L. Shan, Association between insulin receptor gene polymorphism and the metabolic syndrome in Han and Yi Chinese, Asia Pac. J. Clin. Nutr. 21 (2012) 457–463.