Prediction model of random forest for the risk of hyperuricemia in a Chinese basic health checkup test

RF and hyperuricemia

Authors information (in order):

Yuhan Gao 1,2, Shichong Jia 3, Dihua Li 4, Chao Huang 5, Zhaowei Meng 1, Yan Wang 6, Mei Yu 7, Tianyi Xu 7, Ming Liu 8, Jinhong Sun 9, Qiyu Jia 9, Qing Zhang 9, Ying Gao 9, Kun Song 9, Xing Wang 9, Yaguang Fan 10

Affiliations of Authors (in order):

1. Department of Nuclear Medicine, Tianjin Medical University General Hospital, Tianjin, P.R.China
2. School of Medical Imaging, Tianjin Medical University, Tianjin, P.R.China
3. School of Medicine, Shanghai Jiao Tong University, Shanghai, P.R.China
4. Tianjin Key Laboratory of Acute Abdomen Disease Associated Organ Injury and ITCWM Repair, Institute of Acute Abdominal Diseases, Tianjin Nankai Hospital, Tianjin, China
5. Senior Lecturer in Statistics, Hull York Medical School, University of Hull, Hull, UK
6. Tianjin University of Traditional Chinese Medicine, Jian Kang Chan Ye Yuan, Jinghai District, Tianjin, PR China
7. College of Intelligence and Computing, Tianjin Key Laboratory of Advanced Networking (TANK Lab), Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China
8. Department of Endocrinology and Metabolism, Tianjin Medical University General Hospital,
Tianjin, P.R.China

9. Department of Health Management, Tianjin Medical University General Hospital, Tianjin, P.R.China

10. Tianjin Key Laboratory of Lung Cancer Metastasis and Tumor Microenvironment, Tianjin Lung Cancer Institute, Tianjin Medical University General Hospital, Tianjin, PR China

Yuhan Gao, Shichong Jia and Dihua Li as co-first authors contributed equally in the paper.

Correspondence to:

Zhaowei Meng MD PhD, Department of Nuclear Medicine, Tianjin Medical University General Hospital, Anshan Road No. 154, Heping District, Tianjin, P.R.China, 300052 Telephone: 86-18622035159 Fax: 86-022-27813550

E-mail: zmeng@tmu.edu.cn or jamesmencius@163.com

OR

Yan Wang PhD, Tianjin University of Traditional Chinese Medicine, Jian Kang Chan Ye Yuan, Jinghai District, Tianjin 301617, PR China, Telephone: 86-13323497398 Email: wangyan@tjutcm.edu.cn

OR

Mei Yu PhD, College of Intelligence and Computing, Tianjin Key Laboratory of Advanced Networking (TANK Lab), Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China, Telephone: 86-13920397855 Email: yumei@tju.edu.cn

Declarations
Ethics approval and consent to participate

This study was approved by the Institutional Review Board and Ethics Committee of Tianjin Medical University General Hospital and the approval number was 2011-6-1. The study obtained the written informed consent of all participants. We confirm that all methods used in this study were performed in accordance with the relevant guidelines and regulations.

Consent for publication

Not applicable.

Availability of data and materials

The datasets during and/or analyzed during the current study available from the corresponding author on reasonable request.

Competing interests

We declare no competing interests in the paper.

Funding

This study was also supported by the following grants. China National Natural Science Foundation grant #81571709 and #81971650 (to Zhaowei Meng), #81870533 (to Jinhong Sun), #71804124 (to Ying Gao). Key Project of Tianjin Science and Technology Committee Foundation grant #16JCZDJC34300 (to Zhaowei Meng). Tianjin Science and Technology Committee Foundation grant, #12ZCZDSY20400 (to Qiyu Jia), #11ZCGYSY05700 (to Qing Zhang), #13ZCZDSY20200 (to Kun Song) and #17JCYBJC25400 (to Yaguang Fan).

Authors’ contributions

YG and SJ were major contributors in writing the original draft and completed the investigation and formal analysis of the article. ZM, LD, YW participated in the conceptualization,
supervision, review and editing of the article. MY, TX and YF provided algorithm and software support and participated in the experimental design. CH and ML participated in supervision of the article. JS, QJ, QZ, YG, KS, XW provided the data of the article. All authors read and approved the final manuscript.

Acknowledgements

This study was supported by the National Key Clinical Specialty Project (awarded to the Departments of Nuclear Medicine and Radiology).

This study was supported by Tianjin Medical University General Hospital New Century Excellent Talent Program; Young and Middle-aged Innovative Talent Training Program from Tianjin Education Committee; and Talent Fostering Program (the 131 Project) from Tianjin Education Committee, Tianjin Human Resources and Social Security Bureau (to Zhaowei Meng).
Abstract:

Objectives: The present study aimed to develop a random forest (RF) based prediction model for hyperuricemia (HUA) and compare its performance with the conventional logistic regression (LR) model. Methods: This cross-sectional study recruited 91690 participants (14032 with HUA, 77658 without HUA). We constructed a RF based prediction model in the training sets and evaluated it in the validation sets. Performance of the RF model was compared with the LR model by receiver operating characteristic (ROC) curve analysis. Results: The sensitivity and specificity of the RF model were 0.702 and 0.650 in males, 0.767 and 0.721 in females. The positive predictive value (PPV) and negative predictive value (NPV) were 0.372 and 0.881 in males, 0.159 and 0.978 in females. AUC of the RF models was 0.739(0.728-0.750) in males and 0.818(0.799-0.837) in females. AUC of the LR models were 0.730(0.718-0.741) for males and 0.815(0.795-0.835) for females (Figure 1). The predictive power of RF was slightly higher than that of LR, but it was not statistically significant in females (Delong tests, p=0.0015 for males, p=0.5415 for females). Conclusion: Compared with the LR, the good performance in HUA status prediction and the tolerance of features associations or interactions showed great potential of RF in further application. A prospective cohort is necessary for HUA developing prediction. People with high risk factors should be encouraged to actively control to reduce the probability of developing HUA.

Keywords: Hyperuricemia (HUA), random forest (RF), prediction model

Introduction:

Uric acid (UA) is the final metabolite of purines in human. Hyperuricemia(HUA) may result from under-excretion or over-production of UA in congenital or acquired ways[1]. According to the National Health and Nutrition Examination Survey (NHANES) 2015-2016 in the USA, the prevalence of HUA was 20.1% (20.2% and 20.0% in males and females) and was stable during the last decade(p=0.24)[2]. A meta-analysis conducted in China, which pooled study results from 2000 to 2014, showed the prevalence of HUA in China was 13.3% (19.4% and 7.9% in males and females)[3]. A higher prevalence of 16.4% was reported in a separate meta-analysis pooling study from 2000 to 2019 in China (20.4% and 9.8% in males and females)[4]. Although it is lower than
the USA, the prevalence has climbed steadily to over two times during the last two decades and become another common metabolic disease after diabetes mellitus (DM).

Extensive evidence has shown that, in addition to triggering of gout, elevated serum UA is an independent risk factor for chronic kidney disease, hypertension, cardiovascular diseases, dyslipidemia and impaired glucose metabolism, as it also plays an important role in the premature mortality observed in the diseases[5, 6]. It has also been demonstrated that patients with asymptomatic HUA may have urate deposition over joints or even bone erosion, suggesting that asymptomatic HUA to gout is a continuous pathological process[7, 8]. Therefore, it is of great significance to predict HUA early and conduct secondary prevention for high-risk groups.

Thus far, there are few prediction models for HUA. Cao et al developed a Cox regression model using routine anthropometric and blood biomarkers in urban Han Chinese adult[9]. Zeng et al developed an artificial neural network prediction model in Chinese adults based on dietary risk factors[10]. Lee et al tried several machine learning algorithms to predict HUA status in Korea people over 40[11]. As a good clinical prediction model, it should have the characteristics of high prediction power, easy-understanding and convenient-operating. Random forest (RF) is an algorithm that integrates multiple decision trees through the idea of ensemble learning and is capable of representing high order interactions. With further research on HUA, predictors of the model can also be supplemented and replaced without worrying about interaction or association between variables. In this study, based on the RF model, we established a gender-specific prediction model for HUA, and compared its performance with the conventional logistic regression (LR) model, with the aim of developing a predictive model that can be easily generalized and preventing more further adverse health consequences from HUA.

Methods

Study population

The study protocol was approved by the Institutional Review Board and Ethics Committee of Tianjin Medical University General Hospital, China (IRB number 2011-6-1). We analyzed a previously constructed database from a basic health check program during September 2011 to April 2014. 148210 subjects underwent the health check program. Cases with any vacant data were deleted, and a total of 91,690 (83.56%) subjects were invited into this cross-sectional study.
Measurements and definitions

Anthropometric measurements and blood tests of the participants were performed during their visits to our institutions. Height and weight were measured in centimeters and kilograms. Body mass index (BMI) was calculated by dividing weight (kg) by the square of height (m^2). Waist circumference (WC) was measured in centimeters. Systolic and diastolic blood pressures (SBP/DBP) were measured by an automated sphygmomanometer while the subjects were in a seated position after resting for 5 minutes. Blood tests were measured by an auto-analyzer (Hitachi Model 7600 analyzer, Hitachi, Tokyo, Japan). Alcohol intake and smoking history were from the self-recorded questionnaires.

The diagnostic criteria for HUA was SUA>420μmol/L in males and >360μmol/L in females[12]. The diagnostic criteria for DM was fasting plasma glucose ≥7.0mmol/L[13].

Statistical analysis

We tested all parameters for the normality by the Kolmogorov-Smirnov test. The subject characteristics were assessed using Mann-Whitney U test. Differences were considered significant at P<0.05. The results were represented as median (quartile 1-quartile 3), otherwise stated. Odds ratio (OR) with 95% confidence interval (CI) was calculated by the LR. Statistical analysis was performed using SPSS version 26 (IBM Corporation). Comparison of the receiver operating characteristic (ROC) curves was done with the Delong test using MedCalc version 19.

RF model

In this study, tree parameters trained in male dataset included: 'criterion' as 'entropy', 'max_depth' as 10, 'min_samples_leaf': as 12 and 'n_estimators' as 70; and in female dataset included: 'criterion' as 'gini', 'max_depth' as 10, 'min_samples_leaf' as 12 and 'n_estimators' as 90. Models were developed using Python 3.7.

Model evaluation

The discriminatory power of models was analyzed by ROC curves. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated. Cutoff score with the maximum sum of sensitivity and specificity was considered optimum[14].

Results

Studying characteristics
A total of 91690 subjects were included in this cross-sectional study. The cohort was randomly divided into a training set (ratio of 0.8, n=73351) and a validation set (ratio of 0.2, n=18339). Since the number of non-HUA was much larger than HUA, simple down-sampling was carried out on the non-HUA set based on the sample size of HUA to adjust for data imbalance. As a result, the sample distribution of the training sets was 9077/8986 for male and 2054/2033 for female (HUA/non-HUA). The validation set retained the real-world population distribution. The distribution of the validation sets was 2399/8123 (29.53%) for male and 502/7315 (6.86%) for female (HUA/non-HUA).

A total of 21 variables including anthropometric measurements, blood tests, alcohol intake and smoking history were firstly examined in the RF model (shown in the supplementary material), of which the top 10 with the highest weight were selected for each gender (Table 2). All variables differed significantly between groups (p<0.001). 8/10 variables were consistent across gender models. Among HUA patients, the distribution of age and TC were higher in female (p<0.001), and the other variables were significantly higher in males (p<0.001).

**Evaluation of RF model predictive ability**

Classification matrix of the validation set was shown in table 1. The sensitivity and specificity of the RF model were 0.702 and 0.650 in males, 0.767 and 0.721 in females. PPV and NPV were 0.372 and 0.881 in males, 0.159 and 0.978 in females. Area under the curve (AUC) of the RF models was 0.739 (0.728–0.750) in males and 0.818 (0.799–0.837) in females.

**Evaluation of LR model predictive ability**

We further incorporated these parameters into the conventional LR model. First, we conducted univariate LR in both gender groups and all variables except FPG showed statistical significance (Table 3). Next, we conducted multiple LR for both. Since LR model is sensitive to the collinearity of variables, we screened the 10 variables in model 1 with tolerance=0.2 as the cutoff in each gender. For males, BMI, weight and WC suggested significant collinearity, so that we only left one. Multiple LR suggested that variables except WBC were of statistical significance in BMI-retained model (model 2). Interestingly, all variables were shown to be statistically significant when weight was the one to be retained (not shown in the table). For females, collinearity was shown between weight and BMI. The BMI-retained model suggested that WC was not statistically significant, while weight-retained model showed a contradictory result. AUC
of the LR models were 0.730(0.718-0.741) for males and 0.815(0.795-0.835) for females (Figure 1). The predictive power of RF was slightly higher than that of LR, although it was not statistically significant in females (Delong tests, p=0.0015 for males, p=0.5415 for females).

**Discussion**

In the study, we obtained a balanced training set by simple down-sampling, then developed a gender-specific RF model to predict the status of HUA and compared its performance with the most commonly used LR model. It showed that the predictive power of RF was higher than LR in both genders, although it was not significant in females.

Frequently the classifiers based on medical data are developed using class-imbalanced data, where the number of with a particular disease is far less than the healthy. Our study contained 91690 participants, of which 14032 were HUA and 77658 were non-HUA. Standard classification methods applied to class-imbalanced data tend to produce classifiers biased towards the majority class[15]. Therefore, increasing attention has been paid to between-class imbalance problem to improve the accuracy of the minority class prediction[16]. However, practice of imbalanced data processing is not widely used in clinical practice[17]. In this study, based on our relatively sufficient sample size, we tried simple down-sampling to ameliorate the data imbalance[17].

Machine learning has been very popular in various fields for years. Due to its complexity in understanding and operation, it has not been fully applied in clinical work. RF, as an ensemble learning algorithm, integrates multiple decision trees, of which each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest[18]. Compared with the conventional statistical methods, it has the following advantages: 1. RF is a high-throughput algorithm that can handle high-dimensional features without dimension reduction. 2. RF gives the weighted value of each feature in the classification, so that the impact of variables on the classification outcome can be quantified and comparable. 3. RF can deal with imbalanced or missing data to some extent. 4. Interaction or association between variables does not impact much on the classification, as RF can represent high-order interaction[19].

In this paper, we used two different methods as RF and LR to predict the status of HUA. The variables based on RF were significantly different between HUA and non-HUA (Table 2),

---

**Discussion**

In the study, we obtained a balanced training set by simple down-sampling, then developed a gender-specific RF model to predict the status of HUA and compared its performance with the most commonly used LR model. It showed that the predictive power of RF was higher than LR in both genders, although it was not significant in females (Delong tests, p=0.0015 for males, p=0.5415 for females).

Frequently the classifiers based on medical data are developed using class-imbalanced data, where the number of with a particular disease is far less than the healthy. Our study contained 91690 participants, of which 14032 were HUA and 77658 were non-HUA. Standard classification methods applied to class-imbalanced data tend to produce classifiers biased towards the majority class[15]. Therefore, increasing attention has been paid to between-class imbalance problem to improve the accuracy of the minority class prediction[16]. However, practice of imbalanced data processing is not widely used in clinical practice[17]. In this study, based on our relatively sufficient sample size, we tried simple down-sampling to ameliorate the data imbalance[17].

Machine learning has been very popular in various fields for years. Due to its complexity in understanding and operation, it has not been fully applied in clinical work. RF, as an ensemble learning algorithm, integrates multiple decision trees, of which each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest[18]. Compared with the conventional statistical methods, it has the following advantages: 1. RF is a high-throughput algorithm that can handle high-dimensional features without dimension reduction. 2. RF gives the weighted value of each feature in the classification, so that the impact of variables on the classification outcome can be quantified and comparable. 3. RF can deal with imbalanced or missing data to some extent. 4. Interaction or association between variables does not impact much on the classification, as RF can represent high-order interaction[19].

In this paper, we used two different methods as RF and LR to predict the status of HUA. The variables based on RF were significantly different between HUA and non-HUA (Table 2),
which strengthened the theoretical basis and practical confidence for further HUA developing prediction model. The model developed by RF showed high value in predicting HUA. AUC of the RF models was 0.739(0.728-0.750) in males and 0.818(0.799-0.837) in females. Note that the PPV of both gender groups are not high, which may be related to the prevalence of HUA in the population. When sensitivity and specificity in a diagnostic test remained unchanged, higher PPV will be obtained in population with a higher prevalence. However, due to the relatively low prevalence in the real-world data, the true positive sets of the diagnostic test do not increase significantly even if the sensitivity increases, so the method based on adjusting sensitivity has little effect on improving the positive predictive value. By increasing specificity can significantly reduce the number of false positive sets, so as to improve the PPV. However, the model is for screening HUA in the population, so that the sensitivity should not be abandoned while the specificity is improved. Therefore, we made a trade-off in our study and chose the cutoff point with the maximum sum of sensitivity and specificity.

Comparing with LR, we found that the AUC of RF was higher in both gender models, although it was not significant in females. However, as mentioned above, any interaction or association between variables does not adversely affect the RF classification while it does in LR. Since LR is easily affected by the collinearity of variables, when building the LR model, we took the tolerance=0.2 as the bound value, and only one variable in the collinearity set is reserved. We found that when retaining different variables, the effect on the remaining variables was different. Indeed, when one confounding factor is adjusted for, any relevant variable will also be changed. In this case, the choice about the retained variables becomes difficult, especially when there are many variables. At this point, it is methodologically better to use RF, which that is not that sensitive to collinearity. To sum up, we believe that RF has great potential in HUA prediction.

As for the risk factors, in addition to indicators discussed in previous studies, we found that the role of FPG is worth mentioning. FPG ranked 10th in males and 15th in females. After adjusting other covariates, there was no significant association between FPG level and HUA in normal FPG group (P=0.154), but a negative association was found in DM group in males (OR=0.811, P=0.002), which was in consistent with previous studies[20]. Previous studies
have indicated that diabetic patients with glycosuria had a null prevalence of HUA and excreted more UA than those without glycosuria[21]. Chino et al. showed that glycosuria induced by SGLT2 inhibitors at the proximal tubule may inhibit UA reabsorption[21]. Glycosuria resulting from elevated glucose levels in diabetic patients might lead to a competitive inhibition of UA reabsorption. However, the relationship between FPG and serum UA still remains controversial. Studies have reported that serum UA can cause pancreatic β-cell dysfunction, and HUA become an independent risk factor for type 2 DM[22, 23]. In addition, the weight of FPG in female classification is relatively low, and whether there is an interaction between FPG and hormones still needs to be further explored.

The study also has some limitations. First, the dataset is based on a single-center cross-sectional study, and there may be selection bias. Secondly, underlying disease of participants from the program could not be as detailed as the hospital records, which might introduce confounding factors.

Conclusions:

In conclusion, we developed a prediction model of HUA status based on a cross-sectional data, and the model has achieved good results. Compared with the LR, RF has its own irreplaceable advantages, and we believe that RF has great potential in HUA prediction. A further prospective cohort is necessary for HUA developing prediction. Moreover, variables in this study only included common anthropometric and blood biomarkers. Other variables such as lifestyle, education background and income may also be integrated as a comprehensive epidemiological prediction model. This study is only based on the preliminary study of machine learning. We believe that it will provide a direction for the further study of machine learning in clinical research.

Reference

1. Ghei M, Mihailescu M, Levinson D: Pathogenesis of hyperuricemia: recent advances. Curr Rheumatol Rep 2002, 4(3):270-274.
2. Chen-Xu M, Yokose C, Rai SK, Pillinger MH, Choi HK: Contemporary Prevalence of Gout and Hyperuricemia in the United States and Decadal Trends: The National Health and Nutrition Examination Survey, 2007-2016. Arthritis Rheumatol 2019, 71(6):991-999.
3. Liu R, Han C, Wu D, Xia X, Gu J, Guan H, Shan Z, Teng W: Prevalence of Hyperuricemia and Gout in Mainland China from 2000 to 2014: A Systematic Review and Meta-Analysis. Biomed Res Int 2015, 2015:762820.
4. Li Y, Shen Z, Zhu B, Zhang H, Zhang X, Ding X: Demographic, regional and temporal trends of hyperuricemia epidemics in mainland China from 2000 to 2019: a systematic review and meta-analysis. *Glob Health Action* 2021, 14(1):1874652.

5. Bardin T, Richette P: Impact of comorbidities on gout and hyperuricaemia: an update on prevalence and treatment options. *BMC Med* 2017, 15(1):123.

6. Wu AH, Gladden JD, Ahmed M, Ahmed A, Filippatos G: Relation of serum uric acid to cardiovascular disease. *Int J Cardiol* 2016, 213:4-7.

7. Dalbeth N, House ME, Aati O, Tan P, Franklin C, Horne A, Gamble GD, Stamp LK, Doyle AJ, McQueen FM: Urate crystal deposition in asymptomatic hyperuricaemia and symptomatic gout: a dual energy CT study. *Ann Rheum Dis* 2015, 74(5):908-911.

8. Puig JG, Beltran LM, Mejia-Chew C, Tevar D, Torres RJ: Ultrasonography in the diagnosis of asymptomatic hyperuricemia and gout. *Nucleosides Nucleotides Nucleic Acids* 2016, 35(10-12):517-523.

9. Cao J, Wang C, Zhang G, Ji X, Liu Y, Sun X, Yuan Z, Jiang Z, Xue F: Incidence and Simple Prediction Model of Hyperuricemia for Urban Han Chinese Adults: A Prospective Cohort Study. *Int J Environ Res Public Health* 2017, 14(1).

10. Zeng J, Zhang J, Li Z, Li T, Li G: Prediction model of artificial neural network for the risk of hyperuricemia incorporating dietary risk factors in a Chinese adult study. *Food Nutr Res* 2020, 64.

11. Lee S, Choe EK, Park B: Exploration of Machine Learning for Hyperuricemia Prediction Models Based on Basic Health Checkup Tests. *J Clin Med* 2019, 8(2).

12. Zhang Q, Lou SS, Meng ZW, Ren XJ: Gender and age impacts on the correlations between hyperuricemia and metabolic syndrome in Chinese. *Clin Rheumatol* 2011, 30(6):777-787.

13. American Diabetes A: 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2021. *Diabetes Care* 2021, 44(Suppl 1):S15-S33.

14. Liu J, Tang ZH, Zeng F, Li Z, Zhou L: Artificial neural network models for prediction of cardiovascular autonomic dysfunction in general Chinese population. *BMC Med Inform Decis Mak* 2013, 13:80.

15. Nathalie J, Shaju S: The class imbalance problem: A systematic study. *Intelligent Data Analysis* 2002, 6(5):429-449.

16. Haibo H, Garcia EA: Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 2009, 21(9):1263-1284.

17. Rok B, Lara L: Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics* 2010, 11:S23.

18. Breiman L: Random Forests. *Machine Learning* 2001:5-32.

19. Sugimoto K, Murata H, Hirasa H, Aihiara M, Mayama C, Asaoka R: Cross-sectional study: Does combining optical coherence tomography measurements using the ‘Random Forest’ decision tree classifier improve the prediction of the presence of perimetric deterioration in glaucoma suspects? *BMJ Open* 2013, 3(10):e003114.

20. Kuo KT, Chang YF, Wu IH, Lu FH, Yang YC, Wu JS, Chang CJ: Differences in the association between glycemia and uric acid levels in diabetic and non-diabetic populations. *J Diabetes Complicat* 2019, 33(8):511-515.

21. Andrade JAM, Kang HC, Greffin S, Garcia Rosa ML, Lugon JR: Serum uric acid and disorders of glucose metabolism: the role of glycosuria. *Brazilian Journal of Medical and
22. Ghasemi A: Uric acid-induced pancreatic beta-cell dysfunction. *BMC Endocr Disord* 2021, 21(1):24.

23. Mortada I: Hyperuricemia, Type 2 Diabetes Mellitus, and Hypertension: an Emerging Association. *Curr Hypertens Rep* 2017, 19(9):69.
Figure Legend:

Fig 1. Discriminatory power of the two models

The figures showed the discriminatory power of RF (blue) and LR (red) for HUA (A for males; B for females). AUC of RF was higher in RF in both genders. Delong tests showed the significance in males.
Table 1. Classification matrix in different gender

| Case status | Prediction classification |         |         |       | Prediction classification |         |         |       |
|-------------|---------------------------|---------|---------|------|---------------------------|---------|---------|------|
|             | HUA | Non-HUA | Total  |     | HUA | Non-HUA | Total  |     |
| HUA         | 1685  | 714   | 2399 |     | 385 | 117 | 502 |     |
| Non-HUA     | 2839  | 5284  | 8123 |     | 2040 | 5275 | 7315 |     |
| Total       | 4524  | 5998  | 10522 |     | 2425 | 5392 | 7817 |     |
| Variables | Non-HUA       | HUA           | P value | Weight |
|-----------|---------------|---------------|---------|--------|
| Males     |               |               |         |        |
| TG        | 1.34 (0.95-1.96) | 1.83(1.28-2.63) | <0.001 | 0.16   |
| Cr        | 78.00 (71.00-85.00) | 82.00(75.00-90.00) | <0.001 | 0.14   |
| ALT       | 22.00(17.00-32.00) | 28.00(20.00-43.00) | <0.001 | 0.09   |
| BMI       | 25.30(23.20-27.50) | 26.90(24.90-29.10) | <0.001 | 0.09   |
| Weight    | 74.80(67.60-82.10) | 80.20(73.00-88.40) | <0.001 | 0.08   |
| Age       | 46.00(35.00-57.00) | 41.90(32.00-53.00) | <0.001 | 0.07   |
| WC        | 88.00(82.00-94.00) | 92.00(87.00-98.00) | <0.001 | 0.06   |
| TC        | 5.01(4.42-5.65) | 5.27(4.62-5.92) | <0.001 | 0.05   |
| FPG       | 5.00(4.60-5.50) | 5.10(4.70-5.50) | <0.001 | 0.04   |
| WBC       | 5.70(4.90-6.70) | 6.00(5.20-6.90) | <0.001 | 0.03   |
| Females   |               |               |         |        |
| TG        | 0.95(0.68-1.39) | 1.52(1.04-2.21) | <0.001 | 0.16   |
| Cr        | 59.00(53.00-64.00) | 66.00(58.00-72.25) | <0.001 | 0.13   |
| BMI       | 22.90(20.80-25.40) | 26.20(23.30-29.20) | <0.001 | 0.12   |
| WC        | 76.00(70.00-83.00) | 85.00(77.75-92.25) | <0.001 | 0.08   |
| ALT       | 15.00(12.00-20.00) | 19.00(15.00-28.00) | <0.001 | 0.07   |
| Weight    | 58.90(53.50-65.30) | 65.85(59.00-73.23) | <0.001 | 0.06   |
| BU        | 4.30(3.50-5.00) | 5.00(4.20-5.90) | <0.001 | 0.06   |
| Age       | 43.00(31.00-55.00) | 55.45(34.98-65.83) | <0.001 | 0.05   |
SBP \hspace{1cm} 115.00(105.00-130.00) \hspace{1cm} 130.00(115.00-145.00) \hspace{1cm} <0.001 \hspace{1cm} 0.04

TC \hspace{1cm} 4.99(4.35-5.71) \hspace{1cm} 5.42(4.68-6.21) \hspace{1cm} <0.001 \hspace{1cm} 0.04

TG = triglyceride, Cr = creatinine, ALT = alanine aminotransferase, BMI = body mass index, WC = waist circumference, TC = total cholesterol, FPG = fasting plasma glucose, WBC = white blood cell, SBP = systolic blood pressure. Kolmogorov-Smirnov test showed P values of all data were less than 0.001. Mann-Whitney U test was carried out in all variables. Results were represented as median (the first quartile – the third quartile). Feature value was rounded to two decimal places.
Table 3  Analysis of risk factors on HUA based on logistic regression model

| Sex/Variables | Model 1 |       | Model 2 |       |
|---------------|---------|-------|---------|-------|
|               | Crude OR(95%CI) | P     | Adjusted OR(95%CI)* | P     |
| **Males**     |         |       |         |       |
| TG            | 1.288(1.247-1.330) | <0.001 | 1.207(1.165-1.251) | <0.001 |
| Cr            | 1.037(1.033-1.041) | <0.001 | 1.043(1.038-1.047) | <0.001 |
| ALT           | 1.016(1.014-1.018) | <0.001 | 1.007(1.005-1.009) | <0.001 |
| BMI           | 1.161(1.145-1.177) | <0.001 | 1.136(1.118-1.153) | <0.001 |
| Weight        | 1.046(1.042-1.050) | <0.001 | -       |       |
| Age           | 0.985(0.981-0.988) | <0.001 | 0.981(0.977-0.984) | <0.001 |
| WC            | 1.053(1.047-1.058) | <0.001 | -       |       |
| TC            | 1.228(1.229-1.348) | <0.001 | 1.163(1.102-1.228) | <0.001 |
| FPG           | 0.969(0.932-1.007) | 0.112  | 0.871(0.830-0.913) | <0.001 |
| <7.0          |         |       |         |       |
| ≥7.0          |         |       |         |       |
| WBC           | 1.147(1.112-1.183) | <0.001 | 1.027(0.992-1.064) | 0.135  |
| **Females**   |         |       |         |       |
| TG            | 1.596(1.476-1.726) | <0.001 | 1.302(1.202-1.410) | <0.001 |
| Cr            | 1.068(1.059-1.078) | <0.001 | 1.061(1.051-1.072) | <0.001 |
| BMI           | 1.229(1.201-1.258) | <0.001 | 1.172(1.120-1.227) | <0.001 |
| WC            | 1.076(1.067-1.086) | <0.001 | 1.005(0.986-1.023) | 0.612  |
| ALT           | 1.021(1.016-1.026) | <0.001 | 1.010(1.005-1.015) | <0.001 |
|     | Estimate | SE     | 95% CI          | p-value |
|-----|----------|--------|-----------------|---------|
| Weight | 1.071(1.062-1.080) | <0.001 | -               |         |
| BU    | 1.595(1.493-1.705)  | <0.001 | 1.263(1.164-1.370) | <0.001 |
| Age   | 1.033(1.027-1.039)  | <0.001 | 0.987(0.978-0.995) | 0.002   |
| SBP   | 1.032(1.027-1.036)  | <0.001 | 1.010(1.004-1.016) | 0.001   |
| TC    | 1.452(1.339-1.575)  | <0.001 | 1.062(0.960-1.175) | 0.240   |

*Adjusted with other 9 selected covariates.
