Journal of Beijing Institute of Technology, 2021, Vol. 30, Suppl.1

High-Risk People Identification and Key Risk Factor Comparison of Cardiovascular Diseases in Northern China Rural and Urban Areas

Wenjia Chen[1], Jinlin Li[1,✉], He Wang[1], Ying Gao[2], Shuangquan Xin[2], Jinchuan Cao[3] and Hao Li[1]

(1. School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China; 2. Center for Disease Control and Prevention of Keerqin Youyiqianqi, Xing'an League 137423, Inner Mongolia, China; 3. Center for Disease Control and Prevention of Helingeer, Hohhot 011500, Inner Mongolia, China)

Abstract: Data mining techniques were utilized to improve the regional cardiovascular diseases (CVDs) prevention and in addition to find why rural residents are in worse statues of CVDs. Decision tree (DT) was chosen to build a high-risk CVDs people identification model, and the steps of identifying are reduced, with 99.4% precision in the classification. In comparison with the top 18 key features assessed by Random Forest, rural people were at higher risk due to the factors such as higher total cholesterol, blood glucose, blood pressure, body mass index (BMI) and waistline. Urban people had higher hypertensive historical prevalence but lower hypertensive comorbidity prevalence than rural people because of better control of high blood pressure. Smoking, drinking and medicine taking affected current blood pressure and lipid. More tobacco use of rural female and more alcohol intake of rural male can take part of responsibilities for worse CVDs prevalence and control in rural area. In the intervention and prevention of CVDs in a region, high-risk people should be more focused on. And rural and urban residents should be separately implemented with intervenable methods because of different disease history statues, control effect and current statue. Gender also should be considered when lifestyles are significantly different.

Key words: regional cardiovascular diseases prevention; data mining; high-risk group identification; key risk factors identification; intervenable risk factor analysis

CLC number: TP181 **Document code:** A **Article ID:** 1004-0579(2021)S1-0236-11

In non-communicable diseases, cardiovascular diseases (CVDs) cause the largest numbers of death. Although due to encouraging advancements in the prevention, diagnosis, and treatment of CVDs, death rates decreased, while the total worldwide deaths increased by 21.1% during the ten years, resulting in big financial burdens[1]. Both males and females have experienced increasing trends in age-standardized years lived with disability rates, and males experienced worsen trends than females[2]. In China, the prevalence and mortality of CVDs is continuously rising, about 290 million patients suffer from it now, and the mortality rate in rural areas has exceeded and consistently remained higher than that of urban areas in recent years, which has become the most important public health issue[3].

1 Background

Many studies have worked out the risk

Received 2020-01-17
Supported by Key Program of National Natural Science Foundation of China (71432002) ; General Programs of National Natural Science Foundation of China (71972012)
✉ Author for correspondence, professor
E-mail: jinlinli@bit.edu.cn
DOI: 10.15918/j.jbit1004-0579.20012

factors of CVDs, including blood pressure[4-6], blood lipid[7-8], genes[9], comorbidity[10-13], lifestyle[14-16], social determinants[17]. Existence of effective preventive interventions justifies the assertion that CVDs, like heart disease and stroke, could be virtually prevented and controlled[18]. And major modifiable risk factors contributing to CVDs have been identified[14, 19-20]. Hence, it is vital to implement more targeted prevention and intervention measures.

CVDs prediction models in common use were formulated primarily in white populations. In China, in the present, based on China-PAR Project (Prediction for ASCVD Risk in China), 10-year risks and lifetime risk of atherosclerotic cardiovascular disease were predicted[19, 21]. And integrating both population and individual strategies for disease prevention was proposed to end the epidemics of cardiovascular diseases[22]. The app and website for assessing lifetime risk of atherosclerotic cardiovascular disease were built[23]. However, the mechanism of the risk assessment is unclear and runs in the background. After risk assessment, intervention was personalized but has not been regionalized. CVDs prevention improve efficiency considering large population and wide area of China. Moreover, it is important to know statue of CVDs and behavior difference of CVDs risk factor. Rural residents are facing higher CVDs risk than urban residents[24-28]. Cardiovascular medication utilization and adherence rate are of no consistent differences between rural and urban adults[29-30]. Higher quality evidence is needed to determine if differences truly exist between urban and rural patients in the use of, and adherence to, evidence-based medications.

This study focused on regional CVDs prevention in China. And the disease risk from a more macro perspective was assessed, which considered multiple diseases belonging to CVDs and multiple risk factors causing CVDs. Based on the screening dataset, the high-risk group identification model was improved by using the decision tree (DT). And key risk factors were identified by random forest (RF). Differences of key risk factors in the high-risk group among rural and urban people were analyzed, showing the regional characteristics of CVDs. The collected data consisted of more lifestyle information. Intervenable risk factors of CVDs were compared and advices on personalized intervention were suggested.

2 Data and Methodology

2.1 Data description

The study datasets were collected by the Program of Screening and Intervention Subjects with High Risk Cardiovascular Diseases, which was conducted by the Centers for Disease Control and Prevention of Keerqin Youyiqianqi and Helingeer to study the huge public health threats brought by CVDs since 2014. During 2016–2018, 13 021 adults in Keerqin Youyiqianqi (rural) and 14 738 adults in Helingeer (urban) were investigated. Datasets collected in Keerqin Youyiqianqi and Helingeer were respectively remarked as the rural dataset and urban dataset.

In the datasets, the screening results showed whether each individual was high-risk for CVDs, which was the binary outcome for prediction. The datasets totally have 53 features, including basic personal information, lifestyle information, history of diseases, examinations, current disease information, history and current treatment of medicine.

2.2 Data preprocessing

The data preprocessing mainly involves missing data treatment and nonnumeric result conversion. Missing information were respectively filled in with the modes of each features. In the original dataset, disapproves of questionnaires were marked with "0" and approves of questionnaires were marked with "1". All examination results were numerical. A pure numerical dataset was gained after lipid risk results were translated. Lipid risk includes four levels of

"low", "middle", "high", and "extremely high", which can be respectively transform into "0", "1", "2", and "3". Two datasets had the similar gender ratio (male to female); the rural dataset was 0.68 and the urban dataset was 0.66. Furthermore, high-risk individuals took similar percentages in two datasets, 0.28 in the rural dataset and 0.25 in the urban dataset.

2.3 Classification model

We applied the DT method to build binary classifiers based on features after data preprocessing previously described. To build the classifiers, the pre-processed rural dataset was randomly divided into a training dataset (70%) and a testing dataset (30%) by using stratified sampling. Thereby the high-risk people take up the same percentage both in training set and testing set. We also applied other conventional classification methods to build classifiers, like logistic regression (LR), k-NN, Naive Bayes, support vector machine (SVM), and neural network. Classifier performances were measured by mean results of thirty times experiment of three performance measures:accuracy, recall, precision.

2.4 Feature importance

More features mean more time in classification experiments, more cost in collecting data and more work in analyzing data. Besides, high dimension may reduce the performance of classifiers. We applied random forest (RF), with 10 000 estimators, to assess importance rankings of 53 factors, identifying the key risk factors. The key factor combination was imported into classifiers to further demonstrate their importance.

2.5 Statistical analysis

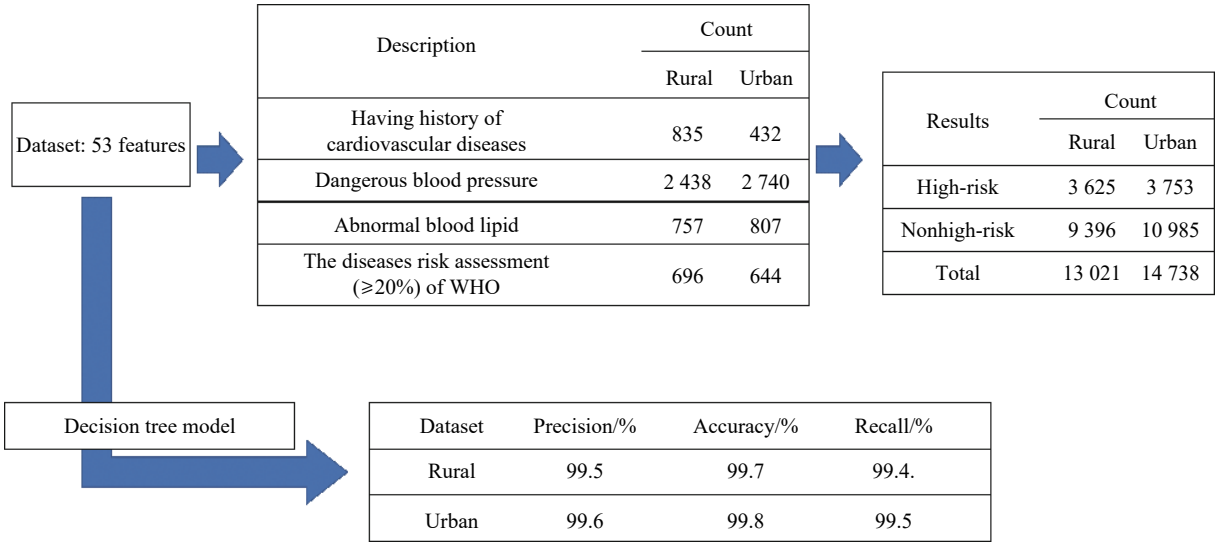
Key risk factors were analyzed by statistical methods. In 19 key risk factors, age and 8 examination results are continuous values while 10 binary factors have results "0" or "1". For continuous-value features, the Shapiro Wilk test was used to assess conformity with a normal distribution. And if the normality is not satisfied, a nonparametric test will be done subsequently. Categorical variables were analyzed using the $\chi 2$ test. In addition to the above key factors, intervenable behaviors, including smoking, drinking, and medicine taking were also analyzed by a mixed linear model. The mixed linear model was used to investigate the effect between three intervenable behaviors and current status of two diseases (hypertensive and dyslipidemia), under adjusting area, age and gender. Due to the significant differences in the lifestyles between females and males, we performed comparison analysis separately for men and women between rural and urban.

3 Results and Discussion

3.1 Improved high-risk group identification model

In the screening program, investigators identified the high-risk people by three steps. As shown in Fig.1, firstly, data of respondents was collected. Secondly, people who had history of cardiovascular diseases, or dangerous blood pressure, or abnormal blood lipid, or the high disease risk assessment (≥20%) of WHO were identified. And then, people who has at least one of the four features in the second step belonged to the high-risk group. In order to reduce the workflow, we used the DT method to build identification model. DT had a high performance on the three measures. In both rural and urban dataset, precision namely high-risk people were correctly identified as high-risk were 99.5% and 99.6%. Accuracy and recall were also as high as precision, ranging in 99.4%–99.8%.

The outperformed classification results were due to the reason that the DT simulated the original identification processing that combines the results of four two-category judgments in Fig.1. Other five classification methods, Naive Bayes, neural network, logistic regression, k-NN, SVM, were also applied to build the high-risk group identification model. Precision increased from 71.4% to 96.3%. Accuracy changed from 83.9% to 92.4%. Recall lied in the range of 65.5% to 84.2%. The performances of those five methods



Description	Count	
	Rural	Urban
Having history of cardiovascular diseases	835	432
Dangerous blood pressure	2 438	2 740
Abnormal blood lipid	757	807
The diseases risk assessment (≥20%) of WHO	696	644

Results	Count	
	Rural	Urban
High-risk	3 625	3 753
Nonhigh-risk	9 396	10 985
Total	13 021	14 738

Dataset	Precision/%	Accuracy/%	Recall/%
Rural	99.5	99.7	99.4.
Urban	99.6	99.8	99.5

Fig. 1 Comparison of original high-risk people identification workflow to DT model

were inferior to that of DT. This is why DT was chosen to build the identification model finally.

3.2 Key risk factor identification

In the well-trained DT model, classification was highly efficient because, in the leaf node of the sixth layer, only 16 of 1 851 high-risk individuals left that needed deeper nodes to identify them. Thus, some features were redundant. Redundant features should be discarded because they do few contributions to the classification, but cause a large structure of the tree. In order to distinguish the contribution to high-risk group identification, the importance of each feature was scored by random forest.

In Tab.1, top 18 features in rural dataset and top 19 features in urban features were ranked from the highest importance score to the lowest score. Those 18 or 19 feature scores summed to 0.95, indicating 95% information was covered.

Mean systolic and diastolic pressures had the highest scores (0.32 and 0.15). Because they were significant and used to diagnose or predict CVDs[4-5], and monitor health statue[31]. Other features had importance scores in the range of 0.008 1–0.076, much lower than that of blood pressure. Using the top important 19 features, performance of DT, with 99.4% precision, was as good as importing all 53 features to identify cardiovascular disease risk. For the identifying high-risk group in this CVDs screening program, those 19 features satisfy the information requirement from the dataset for DT model, so they were key risk factors.

3.3 Key risk factor comparison

We compared 19 key risk factors (Tab. 1) between rural and urban residents. In 19 key risk factors, age and 8 examination results were continuous values and 10 binary factors have results "0" or "1". Tab. 2 presents the comparison of 9 continuous value factors between rural and urban area. Rural female had higher risk for CVDs than urban female because of higher TC, GLU, DBP, and waistline; rural males had higher risk for CVDs than urban male because of higher GLU, SBP, DBP and BMI, and abnormal waistline.

Tab. 3 presents four disease history prevalence. People have much higher hypertensive history prevalence (46.3% – 67.2%) than coronary heart disease (CHD, 4.4% –7.8%), stroke (5.1% – 17.9%), and cardio-cerebrovascular diseases (CCVD, 9.3%–23.2%). Urban people have higher hypertensive historical prevalence than rural people. Both rural and urban females have higher hypertensive historical prevalence than that of males. Urban female has highest hypertensive historical prevalence (67.2%), 19.9% higher than

Tab. 1 Top important features in two datasets

Rank	Rural dataset	Importance	Urban dataset	Importance
1	Mean systolic pressure	0.3193	Mean systolic pressure	0.4146
2	Mean diastolic pressure	0.1440	Mean diastolic pressure	0.1339
3	Cardio-cerebrovascular diseases history	0.0755	High density lipoprotein cholesterol	0.0724
4	Low density lipoprotein cholesterol	0.0710	Low density lipoprotein cholesterol	0.0636
5	High density lipoprotein cholesterol	0.0590	Cardio-cerebrovascular diseases history	0.0428
6	Lipid risk	0.0532	Lipid risk	0.0325
7	Stroke history	0.0430	hypertension	0.0259
8	hypertension	0.0397	Total cholesterol	0.0238
9)	Hypertension aware	0.0274	Hypertension aware	0.0192
10	Hypertension history	0.0213	Hypertension history	0.0175
11	Total cholesterol	0.0213	Stroke history	0.0174
12	Dyslipidemia	0.0147	Triglyceride	0.0135
13	Triglyceride	0.0118	Hypertension control	0.0111
14	Age	0.0108	Age	0.0111
15	Coronary heart disease history	0.0100	Coronary heart disease history	0.0109
16	Blood glucose	0.0098	BMI	0.0105
17	BMI	0.0096	Dyslipidemia	0.0103
18	Waist	0.0081	Blood glucose	0.0090
19			Waist	0.0090
Total		0.9500		0.9500

Tab. 2 Comparison of 9 continuous value factors between rural and urban datasets

Factors	Female		Male		Normal range
	Rural (1978)	Urban (2238)	Rural (1647)	Urban (1515)	
Age	56 (56,8)	61* (60,8)	55 (55,9)	61* (59,10)	-
TC	5.10 (5.25,1.27)	4.89* (5.06,1.20)	4.54 (4.71,1.15)	4.23* (4.34,1.13)	<5.2
LDL	2.82 (2.97,1.07)	2.65* (2.80,1.03)	2.48 (2.60,0.94)	2.27* (2.44,0.96)	<3.4
HDL	1.41 (1.47,0.42)	1.43 (1.47,0.42)	1.25 (1.32,0.48)	1.15* (1.23,0.46)	≥1
TG	1.48 (1.79,1.09)	1.54 (1.81,1.00)	1.39 (1.77,1.18)	1.35 (1.69,1.09)	<1.7
GLU	6.20 (6.64,1.93)	5.90* (6.41,1.91)	6.30 (6.78,1.99)	5.80* (6.27,2.03)	<6.1
SBP	164.5 (160,23.2)	165 (164,19.6)	163.5 (161,22.8)	160.5* (156,20.4)	<140
DBP	92.5 (92,13.0)	91.0* (91,12.0)	97.0 (96,12.9)	91.5* (92,12.6)	<90
Waist	87.0 (87.1,10.76)	85.0* (85.2,10.47)	91.0 (92.0,10.5)	87.0* (87.2,10.3)	F<85, M<90
BMI	26.44 (26.52,3.73)	26.88* (26.90,3.79)	26.51 (25.59,3.47)	26.00* (26.16,3.58)	24, 28

* $P < 0.05$; ** $P < 0.01$; Underlined is out of normal range

Tab. 3 Four disease history prevalence in rural and urban datasets

Disease history	Female		Male	
	Rural (1 978)	Urban (2 238)	Rural (1647)	Urban (1515)
Hypertensive	935 (47.3%)	1504 (67.2%) **	762 (46.3%)	783 (51.7%) *
Coronary heart disease	155 (7.8%)	98 (4.4%) **	96 (5.8%)	112 (7.4%)
Stroke history	313 (15.8%)	115 (5.1%) **	294 (17.9%)	119 (7.9%) **
Cardio-cerebrovascular diseases	458 (23.2%)	208 (9.3%) **	377 (22.9%)	224 (14.8%) **

* $P < 0.05$; ** $P < 0.01$

that of rural female, 20.9% higher than that of urban male, and 15.5% higher than that of urban male. In the rest situation, except that male do not have significant difference in coronary heart

disease prevalence, rural people have 1.55–3.10 times higher historical prevalence than that of urban people. High blood pressure is one of the most important risk factors for some diseases, like CHD, Stroke, CCVD, etc. Therefore, those diseases are always comorbidities of hypertensive.

In order to see the influence of comorbidity, their intersects are listed in Tab. 4. Obviously, rural female has 24.5% lower hypertensive historical prevalence than that of urban female, but has higher historical prevalence of all 6 kinds of comorbidity combinations. Specifically, rural female has 2.12 times higher prevalence of the comorbidity of hypertensive, stroke and CCVD. It can be inferred that rural female did not treat high blood pressure as well as urban female. And in the comorbidity of CHD & CVD,and stroke & CCVD, historical prevalence of rural female is respectively 2.15 and 5.47 times higher than that of urban female, indicating that rural female also did not control some other risk factors of CVDs, except for high blood pressure, as well as urban female. Similarly, rural male has 10% lower hypertensive historical prevalence than that of rural people, but higher historical prevalence of 4 kinds of comorbidity combinations. There was a big gap in comorbidity prevalence of hypertensive, stroke and CCVD (2.50 times higher in rural male), and stroke and CCVD (2.15 times higher in rural male). Urban male has higher historical prevalence of comorbidity of hypertensive, CHD & CCVD and CHD & CCVD, 1.53 and 1.32 times higher respectively. In the past, rural male did not control high blood pressure as well as urban male but control some other risk factors better than urban male.

Tab. 5 presents the current prevalence of hypertensive, and awareness rate of people who have hypertensive, and the rate of people who control hypertensive successfully. Awareness rate (48.7%–69.9%) of people who have hypertensive in those two areas were higher than national level (46.9%), but control rate (2.5%–4.6%) were worryingly lower (15.3%) [6]. Compared to Tab. 3, more (1.37–1.85 times) people have hypertensive. Urban female has the highest current hypertensive prevalence and awareness rate. Prevalence gap between rural and urban female decreased from past 19.9% to current 8.8%. Hypertensive awareness rate of urban female was 19.6% larger than that of rural female. But the control rate did not have significant difference. Thus, hyper-

Tab. 4 Prevalence of comorbidity

Disease history	Female													
	Rural (1 978)							Urban (2 238)						
Hypertensive	√	√	√	√				√	√	√	√			
CHD [a]		√	√		√	√			√	√		√	√	
Stroke		√		√	√		√		√		√	√		√
CCVD [b]		√	√	√	√	√	√		√	√	√	√	√	√
Count	736	5	50	144	5	95	159	1381	3	43	77	2	50	33
Count percent/%	37.2	0.3	2.5	7.3	0.3	4.8	8.0	61.7	0.1	1.9	3.4	0.1	2.2	1.5
Disease history	Male													
	Rural (1 647)							Urban (1 515)						
Hypertensive	√	√	√	√				√	√	√	√			
CHD [a]		√	√		√	√			√	√		√	√	
Stroke		√		√	√		√		√		√	√		√
CCVD [b]		√	√	√	√	√	√		√	√	√	√	√	√
Count	594	5	22	141	8	61	140	698	2	31	52	5	74	60
Count percent/%	36.1	0.3	1.3	8.6	0.5	3.7	8.5	46.1	0.1	2.1	3.4	0.3	4.9	4.0

[a]: Coronary heart disease; [b]: Cardio-cerebrovascular diseases

Tab. 5 Current statue of hypertensive and dyslipidemia

Current statue	Female		Male	
	Rural (1 978)	Urban (2238)	Rural (1647)	Urban (1515)
Hypertensive	1646 (83.2%)	2058 (92.0%) **	1411 (85.7%)	1278 (84.4%)
Hypertensive aware	994 (50.3%)	1565 (69.9%) **	802 (48.7%)	828 (54.7%) **
Hypertensive aware and control	57 (2.9%)	61 (2.7%)	41 (2.5%)	69 (4.6%) *
Dyslipidemia	504 (25.5%)	526 (23.5%)	222 (13.5%)	210 (13.9%)
Dyslipidemia aware	18 (3.6%)	31 (5.9%)	8 (3.6%)	12 (5.7%)
Low lipid risk	684 (34.6%)	674 (30.1%) **	350 (21.3%)	339 (22.4%) **
Middle lipid risk	464 (23.5%)	936 (41.8%) **	338 (20.5%)	354 (23.4%) **
High lipid risk	463 (23.4%)	489 (21.8%) **	678 (41.2%)	662 (43.7%) **
Extremely high lipid risk	367 (18.6%)	139 (6.2%) **	281 (17.1%)	160 (10.6%) **

* $P < 0.05$; ** $P < 0.01$

tensive control is not the main factor for higher comorbidity prevalence of rural female. But for male, current hypertensive prevalence was equal in rural and urban. Urban male has obvious higher awareness rate and control rate, explaining lower prevalence of some comorbidity combinations to some extent. Another important current factor is dyslipidemia. Dyslipidemia prevalence was 13.5%–25.5%, much lower than hypertensive prevalence. And people do not have significantly different prevalence and awareness considering gender and area. But only 3.6%–5.9% people know their dyslipidemia. In lipid risk assessment, rural people were overall more dangerous than urban people, with 3 (female) and 1.61 (male) times more rate of people whose lipid risk was extremely high. It is suggested that dyslipidemia examination and lipid monitoring should be strengthened both in rural and urban, with more focus on monitoring rural people who has extremely high lipid risk.

3.4 Analysis of intervenable risk factors

Smoking, drinking and medicine taking are widely studied as intervenable risk factors of CVDs. In Tab. 6, after adjusted age, gender, and area, non-smoking people was significantly less risky for hypertensive compared to daily smoking people. Seldom smoking and often smoking people's risk of hypertensive were not significant different to daily smoking people. Similarly, non-smoking contributes lower risk to dyslipidemia. Difference of effects to dyslipidemia between seldom or often smoking and daily smoking was non-significant. Namely, smoking is bad for people's blood pressure and lipid no matter how often they smoke.

Tab. 6 Intervenable risk factors of CVDs

Intervenable risk factors		Hypertensive	Dyslipidemia
Smoking	Non	–0.071 (–0.090, –0.051) **	–0.038 (–0.061, –0.014) **
	Seldom	0.039 (–0.008, 0.086)	–0.026 (–0.081, 0.029)
	Often	–0.072 (–0.232, 0.087)	0.159 (–0.029, 0.347)
	Daily	0	0
Drinking	Non	–0.105 (–0.136, –0.075) **	–0.028 (–0.064, 0.007)
	≤ 1 time/month	–0.116 (–0.158,–0.75) **	–0.028 (–0.077,0.021)
	2–4 times/month	–0.081 (–0.112, –0.044) **	–0.021 (–0.064, 0.023)
	2–3 times/week	–0.028 (–0.072, 0.017)	0.004 (–0.048, 0.056)
	≥ 4 times/week	0	0
Regular medication[a]	Yes	–0.174 (–0.190, –0.157) **	–0.028 (–0.047, –0.009) **
	No	0	0

[a]: within the last two weeks;* $P < 0.05$; ** $P < 0.01$

Drinking is another intervenable lifestyle behavior related to CVDs. Compared to more than four drinking per week, non-drinking, no more than one drinking a month and twice to four time a month were significantly much less (0.081–0.116) risk for hypertensive. Controlling alcohol intake is obvious good for the blood pressure. As for blood lipid, how much of alcohol intake has not significantly change the effect to dyslipidemia in this study.

Taking medicine is to control high blood pressure and abnormal blood lipid. Within two weeks, regular medication lead to significantly 0.174 less risk to hypertensive and 0.028 less risk to dyslipidemia. Considering much less people have dyslipidemia and take medicine to control it, smoking, drinking and medicine taking have more obvious (1.87–6.21 times) effects on hypertensive. Their effect to dyslipidemia needs further study based on more data, especially the effect of drinking to dyslipidemia.

Smoking, drinking and medicine taking affect blood pressure and lipid. Those intervenable risk factors were different in rural and urban residents (Tab. 7). Smoking ratio was overall the same between rural and urban area. But considering gender, compared to urban female, the proportion of smoking rural female was 3.69 times larger, and the proportion of everyday smoking rural female was 4.36 times larger. The proportion of smoking male was 2.42–13.00 times larger than female. Compared to rural male, proportion of smoking urban male was 1.36 times larger than that of rural male. Smoking can significantly affect the hypertensive. More use of tobacco of rural female explains low historical hypertensive prevalence but more serious hypertension comorbidity of rural female to some extent. When compared to the tobacco use of nationwide (overall 26%, female 2.1%, male 50.5%, GATS China 2018), the proportion of smoking rural female was 8.09 times larger than the nationwide level; the proportion of smoking urban female and urban male was 2.10 and 1.14 times larger than the nationwide level. Therefore, in this northern China area, control of tobacco use is of great urgency due to its much more use and obvious influence to people's health.

Drinking is an unhealthy lifestyle, which has caused serious damage to Chinese male

Tab. 7 Comparison of intervenable risk factors between rural and urban datasets

Intervenable risk factors	Female		Male	
	Rural	Urban	Rural	Urban
Non-smoking	1641 (83.0%)	2140 (95.6%) **	969 (58.8%)	642 (42.4%) **
Smoking	337 (17%)	98 (4.4%)	678 (41.2%)	873 (57.6%)
Seldom	23 (1.2%)	16 (0.7%)	78 (4.7%)	102 (6.7%)
Often	3 (0.2%)	1 (0.04%)	5 (.3%)	8 (.5%)
Daily	311 (15.7%)	81 (3.6%)	595 (36.1%)	763 (50.4%)
Non-drinking	1715 (86.7%)	2186 (97.7%) **	572 (34.7%)	841 (55.5%) **
Drinking	263 (13.3%)	52 (2.3%)	1075 (65.3%)	674 (44.5%)
≤ 1 time /month	67 (3.4%)	35 (1.6%)	86 (5.2%)	230 (15.2%)
2–4 times / month	115 (5.8%)	7 (0.3%)	269 (16.3%)	198 (13.1%)
2–3 times / week	33 (1.7%)	2 (0.1%)	163 (9.9%)	122 (8.1%)
≥ 4 times /week	48 (2.5%)	8 (0.4%)	557 (33.8%)	124 (8.2%)
Not taking AH[a] drug	300 (30.2%)	129 (8.2%)	293 (36.5%)	120 (14.5%)
Taking AH drug	694 (69.8%)	1436 (91.8%) **	509 (63.5%)	708 (85.5%) **
Can[b]	244 (24.6%)	698 (44.6%)	176 (21.9%)	290 (35.1%)
Cannot[c]	450 (45.2%)	738 (47.2%)	333 (41.5%)	418 (50.5%)

[a]: Antihypertensive; [b]: Can remember the drug name; [c]: Cannot remember the drug name;* $P < 0.05$; ** $P < 0.01$

health[32-33]. Drinking people have 5.78 times larger proportion in rural female (13.3%) than urban female (2.3%). Drinking people take 1.47 times larger percentage in rural male (65.3%) than urban male (44.5%). Overall, rural people have 1.91 times larger proportion of drinking people than urban people. Considering alcohol intake contribute more negative effect to people's hypertensive, alcohol control in this rural area is also important for CVDs prevention and control. More alcohol intake of rural male can also explain the lower historical hypertensive prevalence but higher hypertension comorbidity prevalence to some extent. The risk factors leading to CVDs, like hypertensive and dyslipidemia, are more than smoking and drinking, and complicated affecting and interdependent. In this study, more tobacco use of rural female and more alcohol intake of rural male can take part of the responsibility for more serious situations of CVDs prevalence and worse of CVDs control in rural area. Future intervention of CVDs prevention should focus on this meaningful finding.

Medicine is always taken after the time when people realize they get sick. Female taking antihypertensive medicine and knowing the name of the medicine was 44.3% in urban, 13% more than those (31.3%) in rural, indicating urban female payed more attention on their healthy conditions and try to control the dangerous blood pressures. There was no obvious difference on medicine taking behaviors between rural male and urban male. Totally, urban people did better in antihypertensive medication, explaining the higher historical hypertensive prevalence but lower hypertension comorbidity prevalence compared to rural people on the other aspect. For lipid-lowering medication, because awareness rates of dyslipidemia were low (3.6% –5.9%), even all those people filled their medicine taking information in, there was no significant difference of lipid-lowering medication between rural and urban people. Medicine taking have a significant effect on control of hypertensive and dyslipidemia. In the intervention of antihypertensive medication reminder, rural female should be paid the most attention on. As for lipid-lowering drugs, before reminding medication, making more people know their dyslipidemia should be done firstly.

4 Conclusions

CVDs has become the most important public health issue in China. Before comparing risk factors of CVDs between rural and urban residents, the workflow identifying the high-risk group of CVDs was reduced by using DT technology. The identification model built on the massive screening data collected from 27 759 individuals, which had 99.4% precision in the classification. And then, RF was used to score each feature to differentiate importances of 53 features. Top18 (19) features summed up 0.95 score and contained enough information for high-risk identification, so they were key risk factors. Key factors identification is instructive for future data collection and classifier developing.

In key risk and intervenable risk factors comparison, rural female was of higher risk for CVDs than urban female because of higher TC, GLU, DBP, and waistline; rural males were of higher risk for CVDs than urban male because of higher GLU, SBP, DBP and BMI, and abnormal waistline. People have much higher hypertensive history prevalence (46.3%–67.2%) than coronary heart disease (CHD, 4.4%–7.8%), stroke (5.1%–17.9%), and cardio-cerebrovascular diseases (CCVD, 9.3%–23.2%). Urban people have higher hypertensive historical prevalence but lower hypertensive comorbidity prevalence than rural people. So hypertensive control in rural area is important in this regional CVDs prevention. As for dyslipidemia, only 3.6%–5.9% people know it. It is suggested that both of dyslipidemia examination and lipid monitoring should be strengthened in rural and urban, with more focus on monitoring rural people who has extremely high lipid risk.

Smoking, drinking and medicine taking affect blood pressure and lipid. More tobacco use of rural female and more alcohol intake of rural male can take part of the responsibility for more seriousness of CVDs prevalence and worse situation of CVDs control in rural area. Future intervention of CVDs prevention should focus on this meaningful finding. Totally, urban people did better in antihypertensive medication, explaining the higher historical hypertensive prevalence but lower hypertension comorbidity prevalence compared to rural people on the other aspect. In the intervention of antihypertensive medication reminder, rural female should be paid the most attention on. As for lipid-lowering drugs, before reminding medication, making more people know their dyslipidemia should be done firstly. Knowing people CVDs history statue and current statue in a specific region is important to achieve regional disease intervention and prevention.

References:

[1] James S L G, Abate D, Abate K H, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017 [J]. Lancet, 2018, 392(10159): 1789-1858.

[2] Roth G A, Abate D, Abate K H, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017 [J]. Lancet, 2018, 392(10159): 1736-1788.

[3] Hu Shengshou, Gao Runlin, Liu Lisheng, et al. Summary of the 2018 report on cardiovascular diseases in China [J]. Chinese Circulation Journal, 2019, 34(3): 209-220.

[4] Cho N H, Cho Y R, Park M K, et al. Effect of blood pressure on cardiovascular diseases at 10-year follow-up [J]. The American Journal of Cardiology, 2019, 123(10): 1654-1659.

[5] Parizadeh D, Ghahvehchian H, Asgari S, et al. The association between changes in blood pressure components and incident cardiovascular diseases [J]. Blood Pressure, 2017, 26(6): 341-349.

[6] Wang Z, Chen Z, Zhang L, et al. Status of hypertension in China results from the China hypertension survey, 2012-2015 [J]. Circulation, 2018, 137(22): 2344-2356.

[7] Rai S, Bhatnagar S. Novel lipidomic biomarkers in hyperlipidemia and cardiovascular diseases: An integrative biology analysis [J]. OMICS Journal of Integrative Biology, 2017, 21(3): 132-142.

[8] Zhang M, Deng Q, Wang L, et al. Prevalence of dyslipidemia and achievement of low-density lipoprotein cholesterol targets in Chinese adults: A nationally representative survey of 163641 adults [J]. International Journal of Cardiology, 2018, 267: 218-218.

[9] Folkersen L, Fauman E, Sabater-Lleal M, et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease [J]. Plos Genetics, 2017, 13(4): 1-21.

[10] Upala S, Sanguankeo A, Jaruvongvanich V. Gallstone diseaseand the riskof cardiovascular disease: A systematic reviewand meta-analysis of observational studies [J]. Scandinavian Journal of Surgery, 2017, 106(1): 21-27.

[11] Pavlovic D, Josipovic J, Pavlovic N. Vitamin D in cardiovascularand rental disease prevention [J]. Journal of Medical Biochemistry, 2013, 32(1): 11-15.

[12] Abeles A M. Hyperuricemia, gout, and cardiovascular disease: An update [J]. Current Rheumatology Reports, 2015, 17(3): 495-508.

[13] Bragg F, Holmes M V, Iona A, et al. Association between diabetes and cause-specific mortality in rural and urban areas of China [J]. JAMA Journal of the American Medical Association, 2017, 317(3): 280-289.

[14] Claas S A, Arnett D K. The role of healthy lifestyle in the primordial prevention of cardiovascular disease [J]. Current Cardiology Reports, 2016, 18(6): 1-8.

[15] Frey P, Waters D D. Tobacco smoke and cardiovascular risk: A call for continued efforts to reduce exposure [J]. Current Opinion in Cardiology, 2011, 26(5): 424-428.

[16] Lv J, Yu C, Guo Y, et al. Adherence to healthy lifestyle and cardiovascular diseases in the Chinese population [J]. Journal of the American College of Cardiology, 2017, 69(9): 1116-1125.

[17] Martinez-Garcia M, Salinas-Ortega M, Estrada-Arriaga I, et al. A systematic approach to analyze the social determinants of cardiovascular disease [J].

Plos One, 2018, 13(1): 1-25.

[18] Franco M, Cooper R S, Bilal U, et al. Challenges and opportunities for cardiovascular disease prevention [J]. American Journal of Medicine, 2011, 124(2): 95-102.

[19] Yang X, Li J, Hu D, et al. Predicting the 10-year risks of atherosclerotic cardiovascular disease in Chinese population: The China-PAR project (Prediction for ASCVD Risk in China) [J]. Circulation, 2016, 134(19): 1430-1440.

[20] Dongfeng G, Jiang H, Xiufeng D, et al. Body weight and mortality among men and women in China [J]. JAMA Journal of the American Medical Association, 2006, 295(7): 776-783.

[21] Liu F, Li J, Chen J, et al. Predicting lifetime risk for developing atherosclerotic cardiovascular disease in Chinese population: The China-PAR project [J]. Science Bulletin, 2018, 63(12): 779-787.

[22] Huang M, Liu S. Ending the epidemics of cardiovascular diseases: Time is now to integrate both population and individual strategies for prevention [J]. Science Bulletin, 2018, 63(12): 735-736.

[23] Anonymous. The joint task force for guideline on the assessment and management of cardiovascular risk in China [J]. Chinese Circulation Journal, 2019, 34(1): 4-28.

[24] Harris D E, Hamel L, Aboueissa A M, et al. A cardiovascular disease risk factor screening program designed to reach rural residents of Maine, USA [J]. Rural and Remote Health, 2011, 11(3): 1-15.

[25] Nuotio J, Vahamurto L, Pahkala K, et al. CVD risk factors and surrogate markers-Urban-rural differences[J]. Scandinavian Journal of Public Health, 2019. [published online ahead of print, 2019-08-29]. doi:10.1177/1403494819869816.

[26] Yang M, Lu J, Hao Q, et al. Does residing in urban or rural areas affect the incidence of polypharmacy among older adults in western China? [J]. Archives of Gerontology and Geriatrics, 2015, 60(2): 322-333.

[27] Trivedi T, Liu J, Probst J C, et al. The metabolic syndrome: Are rural residents at increased risk? [J]. Journal of Rural Health, 2013, 29(2): 188-197.

[28] Li W, Song F, Wang X, et al. Prevalence of metabolic syndrome among middle-aged and elderly adults in China: Current status and temporal trends [J]. Annals of Medicine, 2018, 50(4): 345-353.

[29] Murphy G K, McAlister F A, Weir D L, et al. Cardiovascular medication utilization and adherence among adults living in rural and urban areas: A systematic review and meta-analysis [J]. BMC Public Health, 2014, 14: 544.

[30] Murphy G K, McAlister F A, Eurich D T. Cardiovascular medication utilization and adherence among heart failure patients in rural and urban areas: A retrospective cohort study [J]. Canadian Journal of Cardiology, 2015, 31(3): 341-347.

[31] Kario K, Kanegae H, Tomitani N, et al. Nighttime blood pressure measured by home blood pressure monitoring as an independent predictor of cardiovascular events in general practice [J]. Hypertension, 2019, 73(6): 1240-1248.

[32] Griswold M G, Fullman N, Hawley C, et al. Alcohol use and burden for 195 countries and territories, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016 [J]. Lancet, 2018, 392(10152): 1015-1035.

[33] Degenhardt L, Charlson F, Ferrari A, et al. The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016 [J]. Lancet Psychiatry, 2018, 5(12): 987-1012.