The Logistic Regression Analysis on Risk Factors of Hypertension among Peasants in East China & Its Results Validating

Zhuoyuan ZHENG[1] , Ye LI[1] * and Yunpeng CAI[1]

[1] Shenzhen Institutes of Advanced Technology, Key Laboratory for Health informatics, Chinese Academy of Sciences Shenzhen, 518055, China

Abstract

Hypertension can lead to coronary heart disease, heart failure, stroke, kidney failure, and other health problems. It is of great significance for researchers to find out the risk factors that affect the onset, maintenance and prognosis of hypertension. In this study we aim to find out the risk factors of hypertension from questionnaire data for peasants in East China. We develop some customized preprocessing methods according to the characteristics of questionnaire data, employ logistic regression to discover some results and propose a method to validate these results. The final results show that age, FPG, the number of people for dinner, HC, the times of eating pickles are more than 7 per week, drinking, illiterate, moderate labor, drinking groundwater, times of bearing and smoking are risk factors of HBP for peasants in East China.

Keywords: *Logistic Regression, Hypertension, Risk Factors, Questionnaire*

1. Introduction

Hypertension is also called "High Blood Pressure (HBP)". It can lead to coronary heart disease, heart failure, stroke, kidney failure, and other health problems. In china, every year about 80 percent of deaths are attributable to chronic diseases mostly induced by hypertension. The Chinese incidence of hypertension has almost doubled in past 20 years. Peasants comprise 70 percent of the Chinese population. Due to different reasons such as economy, condition, regime and knowledge, hypertension in rural district of China characterize by three high(high prevalence rate, high mortality and high disability rate)and three low(low realizing rate, low treatment rate and low control rate). This situation issues a grim challenge to Chinese medical workers. The good solution of hypertension problems will be of great importance and significance in economy and society.

According to WHL and WHO, the most effective way to tackle hypertension problem is to prevent it before it come into being. So to find out the risk factors that may induce hypertension should be the first task we must carry out. A lot of researchers had finished plenty of effective works in hypertension risk factors study field. Judith E.Neter et al. [1] performed meta-analysis of randomized controlled trials to estimate the effect of weight reduction on blood pressure overall and in population subgroups. The result shows that weight loss is important for the prevention and treatment of hypertension. Peretz Lavie et al. [2] assess whether sleep apnoea syndrome is an independent risk factor for hypertension. Multiple logistic regression showed that each additional apnoeic event per hour of sleep increased the odds of hypertension by about 1%. Paul K.Whelton et al. [3] employ randomized controlled trial to determine whether weight loss or reduced sodium intake is effective in the treatment of older persons with hypertension. They found that reduced sodium intake and weight loss constitute a feasible, effective, and safe non-pharmacologic therapy of hypertension for older persons. John P. Forman et al. [4] estimate hypertension incidence associated with dietary and lifestyle factors in women. Results showed that adopting low-risk dietary and lifestyle factors has the potential to prevent a large proportion of new-onset hypertension occurring among young women.

In the fields of research on risk factors of diseases, plenty of data were acquired through questionnaire of random sampling. For example, L.Nikolajsen and N.J.Talley [5, 6] make use of questionnaire to do their research works. Generally, researchers adopt the classic statistical analysis methods [7, 8] to analysis data obtained. The advancement in artificial intelligence area contributed to medical intelligent diagnoses [9]. Logistic regression models have become the most widely used statistical tool for modeling binary response variables and for analyzing case–control data [10].

In our study, logistic regression is employed to analyze the questionnaire data from Pizhou city in Jiangsu Province so as to find out some risk factors of hypertension among local peasants.

The materials of the paper are organized as follows. Section 2 introduces materials and method used to process questionnaire data in the paper. In section 3, logistic

* correspondence author

regression model is used to analyze data and some results are got. Finally, we validate the results and reach a conclusion in section 4.

2. Materials and Methods

A. Subject

The questionnaire data is designed for risk factors investigation of chronic disease. The data involved 6564 residents (3231 males and 3333 females) from 4 towns of Jiangsu Province in East China. Their age is 42.39±14.64, the oldest is 70 and the youngest is 16.

In the questionnaire, information about chronic diseases (high blood pressure, coronary heart disease, stroke, diabetes and malignant tumors) was enquired and included a total of 11 big items (personal information, living condition, health care, chronic history, familial chronic history, smoking, drink, alcohol use, eating habits, daily living & physical exercise, woman menstruation & birth history) and 221 sub-items.

B. Data Extraction

The questionnaire data include three measurements on systolic blood pressure (SBP) and diastolic blood pressure (DBP) respectively. According to Guidelines for Prevention and Treatment of hypertension in China, if three measured values of SBP are higher than 140 mmHg or three measured values of DBP are higher than 90 mmHg for a response, then we can decide that this is a case of hypertension. In order to increase the contrast between hypertension case and normal case, we extract hypertension data that three measured values of SBP are higher than 140 mmHg and three measured values of DBP are higher than 90 mmHg. At the same time, non-hypertension data was extracted on the basis that three measured values of SBP are less than 120 mmHg and three measured values of DBP are less than 80 mmHg. In this way, 1763 records, including 339 hypertension and 1424 non-hypertension records, are extracted. At last, a new field named 'HBP' was added into data. Corresponding to hypertension records, the value of HBP is '1' while non-hypertension records' is '0'.

C. Preprocessing

Real world data are generally incomplete and inconsistent dirty data. Using these raw data, people are always incapable either to carry out data analysis directly or to get a satisfying result. Data preprocessing techniques can improve data quality and reduce processing time.

In consideration of the complexity of questionnaire data, four major preprocessing methods as show below are employed.

1) Data transformation: Most of data from questionnaire are numbers of some options. For example, species of tumors are inquired in the sub-questionnaire of cancer, and these options include stomach cancer =01, esophagus cancer =02, liver cancer =03, lung cancer =04, colorectal cancer =05, breast cancer =06, cervical cancer =07, leukemia cancer =08 and others =88. In order to transform these data to form which suit to be processed, this field must be divided. And every possible value should be a new field. The value of this new field is either "true" or "false" ("T" or "F", "1" or "0"). The value "true" ("T" or "1") means the corresponding option is picked out, and the value "false" ("F" or "0") means the corresponding option is not picked out. For example, a record numbered 50324 in TABLE 1 can be transformed into the record shown in TABLE 2.

2) Data normalization: In questionnaire, the choice of "01" means a positive response while "02" is a negative one. So in the process of data anglicizing, the choices of "01" and "02" must be modified into "true" ("T" or "1") and "false" ("F" or "0") respectively. Some fields have the values of 00, 01… and 09. These values must be changed to 0, 1… and 9. In addition, due to some inputting mistakes, there are some insignificant values such as "0."

Table 1: Original Record

number	species of tumors
50324	01, 07

and ".". These values can be changed to value "0".

3) Blank handling: Because of the carelessness of investigators or typist, there are some missing values in some fields. For different fields, the methods handling missing values are different. For instance, the missing value about age can be filled by average of age, while the missing value like alcohol intake can be filled by zero.

4) Removing void value: Some attributes like number, name, address and phone numbers are insignificant for risk factors analysis and should not be considered. Besides, there is only one value for all records in some fields, and these fields can not show the discrepancy of different records. So these fields should be excluded from our analysis.

D. Feature selection

Feature selection can generate the minimal subsets of attributes for specific application without loss of data value and remove the irrelevant and redundant attributes.

Table 2: Transformed Record

number	Stomach cancer	esophagus cancer	Liver cancer	lung cancer	colorectal cancer	breast cancer	cervical cancer	leukemia cancer	others
50324	T	F	F	F	F	F	T	F	F

What's more, feature selection can also improve the quality of data, speed up the data analysis and bring the results discovered more comprehensible.

After preprocessing, the data extracted include 307 fields (corresponding to variables). Some fields are not statistically significance. Sometimes these fields perhaps disturb statistic analysis and lead to wrong results. On the other hand, if we take these insignificant fields into account, this will increase the complexity and bring about unintelligible result. On account of above reasons, we employ feature selection to get rid of these insignificant fields. In the end, 205 fields are removed and 102 fields are remained.

3. Logistic Regression Analysis

Logistic regression is a useful tool in analytical epidemiology primarily because the method yields a direct estimate of the odds ratio for any regressor variable (risk factor) statistically adjusting for the confounding effect of other regressor variables included in a given analysis. In our study, we adopt binary logistic regression model for risk factor discovery and use SPSS as analysis tool. In SPSS, field HBP act as dependent variable and the rest are covariates. More details about logistic regression model can be found in [11, 12].

Table 3: Selected Results by Logistic Regression

Variable	B	S.E.	Wald	Sig.	Exp(B)
V1	0.05	0.014	12.363	0	1.051
V2	0.646	0.108	35.994	0	1.908
V3	-0.379	0.092	16.945	0	0.684
V4	-1.043	0.284	13.466	0	0.352
V5	-1.073	0.323	11.034	0.001	0.342
V6	0.057	0.019	9.204	0.002	1.059
V7	0.428	0.144	8.849	0.003	1.534
V8	0.797	0.29	7.559	0.006	2.22
V9	0	0	7.664	0.006	1
V10	0.051	0.019	7.015	0.008	1.053
V11	-0.792	0.296	7.147	0.008	0.453
V12	-0.44	0.166	6.989	0.008	0.644
V13	-0.988	0.377	6.861	0.009	0.372
V14	0.077	0.03	6.519	0.011	1.08
V15	-1.043	0.446	5.477	0.019	0.352
V16	-0.806	0.35	5.311	0.021	0.447
V17	0.694	0.3	5.364	0.021	2.002
V18	0.797	0.358	4.96	0.026	2.218
V19	0.181	0.086	4.439	0.035	1.199
V20	0.036	0.018	3.84	0.05	1.036

S.E.: Standard Error, Sig:Significance,P(Sig.) ≤0.05

Table 4: Selected Variables by Logistic Regression and Corresponding Factors

Variable	Corresponding Factor
V1	age
V2	FPG
V3	the number of people for dinner
V4	the 1st eating habit: often eat fat
V5	the 2nd eating habit: often eat lean meat
V6	HC
V7	amounts of salt intake per month
V8	the 3rd eating habit: the times of eating pickles are more than 7 per week
V9	expenditure of diet per month
V10	years of drinking
V11	the 1st educational level: illiterate
V12	amounts of animal oil intake per month
V13	the 4th eating habit: the times of eating pickles are less than 1 per month
V14	amounts of vegetable oil intake per month
V15	the 1st sleep quality: good sleep
V16	the 5th eating habit: the times of eating pickles are 2 or 3 per month
V17	the 3rd manual labor: moderate labor such as installation worker
V18	drinking groundwater
V19	times of bearing
V20	amounts of cigarette smoked per day

FPG: Fasting Plasma Glucose (mol/L)
HC: Hip Circumference

The result output is shown as Table 3 and Table 4 describes the corresponding relationship between variables selected in logistic regression and factors. All regression variables in Table 3 must be in conformity to a principal that its P (sig.) value is not greater than 0.05. By this standard, only 20 of 102 variables are listed in Table 3.

4. Conclusions

For regressor variables in Table 3, we cannot vague decide that all of them are risk factors, because there are

Table 5:Variables with Binary Value Selected in Logistic Regression and Its Odds Ratio

Variable	HBP			non-HBP			Odds Ratio
	Present	Absent	Odds	Present	Absent	Odds	
V4	190	149	1.2752	934	490	1.9061	0.669
V5	77	262	0.2939	393	1031	0.3812	0.771
V8	146	193	0.7565	301	1123	0.268	2.8228
V11	107	232	0.4612	198	1226	0.1615	2.8557
V13	27	312	0.0865	250	1174	0.2129	0.4063
V15	96	243	0.3951	540	884	0.6109	0.6468
V16	28	311	0.09	362	1062	0.3409	0.264
V17	56	283	0.1979	164	1260	0.1302	1.52
V18	79	260	0.3038	243	1181	0.2058	1.4762

The number of HBP cases is 339. The number of non-HBP cases is 1424.
Odds Ratio= (HBP's Odds)/ (non-HBP's Odds)

not reasonable explanations for some variables. Here we propose a method so as to exclude some variables which may lead to paradoxical results in Table 3.

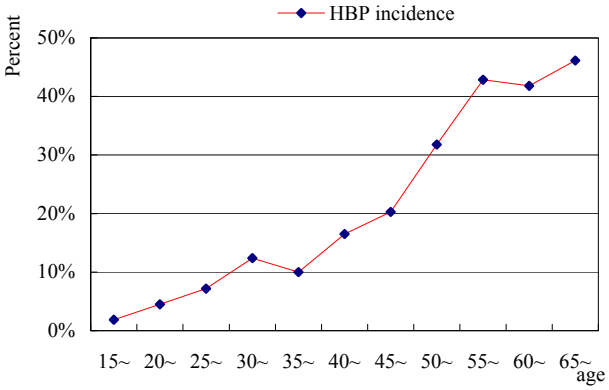

Fig. 1. Change curve of HBP incidence along with the increase of age. We take every 5 years as a stage. E.g. stage 30~ means age is older or equal than 30 and younger than 35.

According to the different data type, variables in Table 3 can be divided into two categories: binary type and numeric type. For variables of binary type (V4, V5, V8, V11, V13, V15, V16, V17, V18), we can calculate their odds ratios and decide which factor is a risk or not according to its corresponding odds ratio. By the interpretation of odds ratio [13], an odds ratio greater than 1 implies a positive association between the factor and

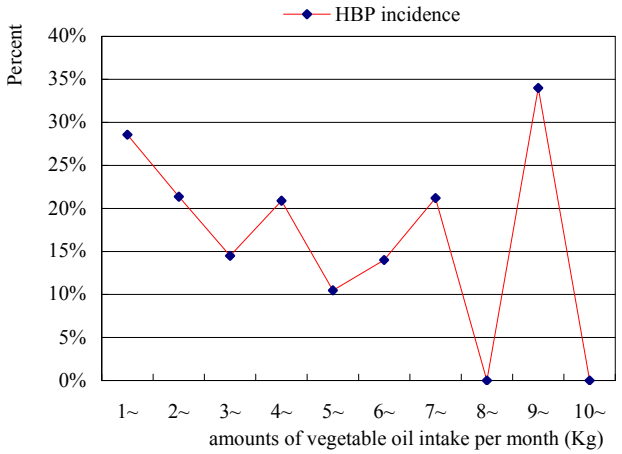

Fig. 2. Change curve of HBP incidence along with the increase of amounts of vegetable oil intake per month.
We take every 1 kg as an interval. E.g. interval 3~ means vegetable oil intake is bigger or equal than 3 kg and less than 4 kg.

hypertension. In Table 5, we list the odds ratio and marked variables with odds ratio less than 1 in red. So we can exclude factors corresponding to variables V4, V5, V13, V15 and V16 as risks. Variables (V8, V11, V17, V18) supporting logistic regression results can be remained and we may take these corresponding factors as risks for hypertension.

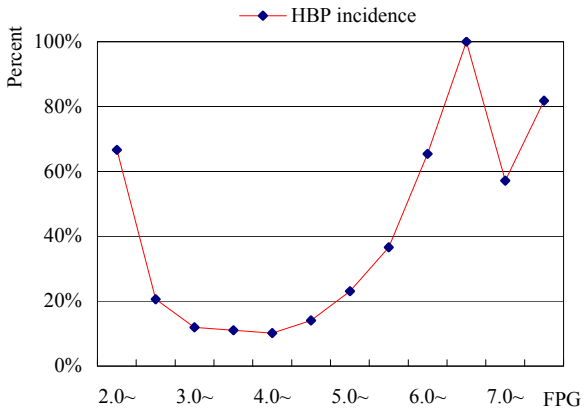

Fig. 3. Change curve of HBP incidence along with the increase of FPG. We take every 0.5 mol/L as an interval. E.g. interval 3.0~ means FPG is higher or equal than 3.0 mol/L and lower than 3.5 mol/L.

As for variables of numeric type (V1, V2, V3, V6, V7, V9, V10, V12, V14, V19, V20), we can not calculate their odds ratios by this way. But we can draw their tendency charts (change curves) which describe the association between these factors and the incidence of hypertension. Based on different characteristics of 11 tendency charts, they fall into three categories. Due to limited space, in each category group only one representative chart is picked out for discussion.

The first category includes variables V1, V3, V6, V10, V19 and V20. The notable feature of this category is that the change curve in each tendency chart shows an upward or downward tendency along with factors' change. Fig.1 is a typical one of this category. From fig.1 we can find that HBP incidence goes up along with the increase of age (V1). Even though at stage 35~ and 60~, the HBP incidence becomes a little decrease comparing with preceding stage, we can still discover the regularity that HBP incidence goes up along with the increase of age on the whole as shown in Fig.1. That suggests that age is a risk factor of HBP. The same goes for the others in this category. So we can conclude that factors corresponding to variables V1, V3, V6, V10, V19 and V20 are risk for HBP.

The second category includes variables V7, V9, V12 and V14. We cannot find out any distinctive characteristic from charts of this category. A representative chart is shown as Fig. 2. HBP incidence in Fig.2 shows irregular change along with the increase of amounts of vegetable oil intake per month. The same goes for the others in this category. So we cannot give a clear conclusion that factors corresponding to variables V7, V9, V12 and V14 are risks or not for HBP.

The third category only includes one variable V2 and the corresponding tendency charts is shown as Fig.3. This chart describes the changing trend of HBP incidence when FPG increase every 0.5 mol/L. From Fig.3 we can find

two changing trend. The first one is that HBP incidence steady decline when FPG increase from 2.0 to about 4.0. The second is that HBP incidence goes up on whole when FPG is greater than 4.0. On the premise of less than 4.0mol/L, the former suggests that the lower FPG is, the higher your risk of HBP. But the latter implies an inverse result. All this shows that too high or too low FPG are risks for HBP.

From what has been discussed above, we may safely draw a conclusion that factors corresponding to variables V1, V2, V3, V6, V8, V10, V11, V17, V18, V19 and V20 (age, FPG, the number of people for dinner, HC, the times of eating pickles are more than 7 per week, drinking, illiterate, moderate labor, drinking groundwater, times of bearing and smoking) are risk factors of HBP for peasants in East China. That will be a good reference for HBP prevention of rural medical work in local district.

Acknowledgments

This work was supported in part by the Comprehensive Strategic Cooperation Projects for Guangdong Province and Chinese Academy of Sciences (Grant No. 2011A090100025), the Promotion and Development Project for Key Laboratory of Shenzhen (Grant No. CXB201104220026A) and the National Natural Science Foundation of Youth Science Foundation (Grant No. 31000447).

References

[1] Judith E. Neter, Bianca E. Stam, Frans J. Kok, Diederick E. Grobbee, Johanna M. Geleijnse, Influence of Weight Reduction on Blood Pressure: A Meta-Analysis of Randomized Controlled Trials, Hypertension 2003,42:878-884, Sep 15.

[2] Peretz Lavie, Paula Herer, Victor Hoffstein, Obstructive sleep apnoea syndrome as a risk factor for hypertension population study, BMJ 2000;320:479-482.

[3] Paul K. Whelton; Lawrence J. Appel; Mark A. Espeland; et al.Sodium Reduction and Weight Loss in the Treatment of Hypertension in Older Persons,JAMA. 1998,279(11):839-846

[4] John P. Forman, Diet and Lifestyle Risk Factors Associated With Incident Hypertension in Women , JAMA. 2009,302(4):401-411

[5] L. Nikolajsen, B. Brandsborg, U. Lucht, T. S. Jensen, H. Kehlet. Chronic pain following total hip arthroplasty: a nationwide questionnaire study [J]. Acta Anaesthesiol Scand 2006; 50: 495-500

[6] N.J.Talley,P.Newman,P.M.Boyce,K.J.Paterson,B.K. Owen. Initial validation of a bowel symptom questionnaire and measurement of chronic gastrointestinal symptoms in Australians [J]. Internal Medicine Journal.Volume 25 Issue 4, pp. 302-308

[7] LF Masson,G MCNeill,JO Tomany,JA Simpson,HS Peace,L Wei,DA Grubb,C Bolton-Smith. Statistical approaches for assessing the relative validity of a food-frequency questionnaire: use of correlation coefficients and the kappa statistic [J]. Public Health Nutrition (2003), 6:313-321

[8] Maria Hagströmer, Pekka Oja,Michael Sjöström. The International Physical Activity Questionnaire (IPAQ): a study of concurrent and construct validity [J]. Public Health Nutrition (2006), 9:755-762

[9] Qeethara Kadhim Al-Shayea, Artificial Neural Networks in Medical Diagnosis, IJCSI International Journal of Computer Science Issues, 2011, 8(2):150-154

[10] N. Breslow, N.E. Day, Statistical Methods in Cancer Research, vol. 1, The Analysis of Case–control Studies, ARC, Lyon, 1980.

[11] Lee J, An Insight on the Use of Multiple Logistic-Regression Analysis to Estimate Association between Risk Factor and Disease Occurrence, International Journal of Epidemiology, MAR 1986, Volume 15, pp: 22-29

[12] Bewick V, Cheek L, Ball J, Statistics review 14: Logistic regression, CRITICAL CARE, 2005,volumn: 9,pp: 112-118

[13] Stuart Spitalnic, Risk Assessment II: Odds Ratio, Hospital Physician, 2006,pp:23-26

Zhuoyuan ZHENG received his B.S. and M.S. degrees of engineering in Computer Application Technology from Guilin University of Electronic Technology, PR China in 2000 and 2003 respectively. Now, he is a PhD candidate in University of Chinese Academy of Sciences, Beijing. His current research interest is medical data mining.

Ye LI is a member of the IEEE. He received the B.S. and M.S. degrees in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering from Arizona State University, Tempe, U.S., in 2006. In 2007, he worked in Cadence Design Systems Inc., San Jose, CA. Since 2008, he has been working as a Professor in Shenzhen Institutes of Advanced Technology (SIAT), Key Lab for Health Informatics, Chinese Academy of Sciences, Shenzhen, China. His research interests include mobile health and tele-healthcare etc. Dr. Li served as the reviewer of several IEEE journal and transactions.

Yunpeng CAI received the PhD degree in computer science from Tsinghua University, Beijing, China, in 2007. He is currently an associate professor with the Center for Biomedical Information Technology, Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen, China. His research interests include bioinformatics, health informatics, machine learning and evolutionary computation.