Dean&Francis
ISSN 2959-6157

An Analysis of the Influencing Factors Associated with Hypertension

Linxi Zhu

School of Information Engineering,
Zhejiang Ocean University,
Zhejiang, 311100, China
gaoqing@asu.edu.pl

Abstract:

In contemporary society, hypertension, as a prevalent cardiovascular disease, has exhibited a steady increase in prevalence and maintains a persistently high mortality rate among cardiac-related conditions. Currently, this disease poses a significant threat to patients' health and well-being. This study aims to investigate the primary factors influencing the incidence of hypertension through an in-depth analysis of demographic data. The research conducted analytical examinations on the official dataset from the Kaggle platform. Initially, descriptive analysis was performed, utilizing histograms to visually demonstrate the impact of factors such as heart rate, cholesterol levels, and Body Mass Index (BMI) index on hypertension. Subsequently, logistic regression analysis was employed to explore the relationships between hypertension incidence and variables including heart rate, cholesterol, age, BMI, and BPMeds. Based on this analysis, a binary logistic regression model was established, and predictive evaluation was conducted using the constructed model. The research findings indicate that the model's overall prediction accuracy is approximately 75%, providing a valuable framework for hypertension risk assessment in both general and specific populations.

Keywords: Hypertension; Binary Logistic Regression; Descriptive statistical analysis

1. Introduction

Hypertension is a common chronic disease. In today's society, the incidence rate and the number of patients with hypertension are constantly increasing. Nowadays, hypertension has become one of the most prevalent chronic non-communicable diseases worldwide, posing a serious public health problem. In recent years, the incidence of hypertension has continued to rise [1]. From 1990 to 2019, the number of global hypertension patients has doubled [2]. Currently, it is estimated that 1.08 million people die from hypertension each year, and more than half of cardiovascular disease deaths worldwide are attributed to hypertension [3]. Therefore, the prevention and treatment of hypertension are of great significance. There are many factors contributing to hypertension, which also makes this disease complex.

ISSN 2959-6157

The distribution of hypertension exhibits a significant age-related pattern, with elderly patients constituting a substantial proportion within the age structure. Hu employed a descriptive statistical approach to categorize the data based on age groups and analyzed the correlation between employees' age and hypertension. The findings revealed a progressive increase in the prevalence of hypertension with advancing age [4]. However, numerous contributing factors persist, including gender, dietary patterns, and exercise habits. Zhang mentioned various influencing factors of hypertension in his report [5]. Additionally, various external determinants continue to influence the onset of hypertension.

Tobacco contains various harmful substances such as nicotine, tar, and carbon monoxide. These substances affect the cardiovascular system through complex physiological mechanisms. Epidemiological and clinical studies have shown that an increase in the amount and duration of smoking can elevate the risk of cardiovascular events [6]. There are still some uncertainties regarding the association between smoking and hypertension, such as the impact of smoking on hypertension in different genders. Conducting in-depth research on the impact of smoking on hypertension can help uncover the onset patterns of cardiovascular diseases like hypertension and provide scientific evidence for hypertension prevention. This study aims to analyze the impact of smoking status on hypertension. This study employs descriptive statistical analysis and constructs a binary linear regression model to investigate the influencing factors of hypertension and make certain predictions.

2. Method

2.1 Data Sources and Notes

The research data were sourced from the initial hypertension dataset compiled by MD Raihan Kahn, which was obtained from Kaggle [7]. Table 1 is derived from the official dataset available on the Kaggle platform, according to the relevant literature. Hypertension is defined as a condition characterized by systolic blood pressure exceeding 130 mmHg and diastolic blood pressure surpassing 80 mmHg [8].

Table 1. Different types of variables

Term	Type	Range
Age	Numeric	31 to 70
Sex	Categorical	0-Female,1-male
Current smoker	Categorical	0-No,1-yes
Heart rate	Numeric	44 to 143
BP	Categorical	0-Not,1-have
Chol	Numeric	113.0 to 696.0
BPMeds	Categorical	0-No,1-yes
BMI	Numeric	15.96 to 56.8

2.2 Method Introduction

Descriptive statistical analysis provides a comprehensive summary and characterization of the data. Through processes such as data collection, organization, and analysis, this methodology enables the computation of key statistical measures. In this study, the distribution and characteristics of the data are visually represented using bar charts as graphical tools, thereby enhancing the understanding of the fundamental attributes and distribution patterns of the data.

Furthermore, this study employs the logit regression methodology to conduct a binary logit regression analysis on the dataset. Logistic regression represents a classical classification model, whose fundamental mechanism involves transforming the linear regression outputs into probability values through a logistic function, thereby enabling the determination of the likelihood of a sample belonging to a specific category. The logistic regression model offers numerous advantages for investigating the determinants of hypertension. Firstly, it exhibits high computational efficiency with a relatively simple model architecture, devoid of intricate parameter optimization processes, resulting in minimal computational load during both training and prediction phases. This enables rapid data processing, making it particularly suitable for large-scale datasets. Secondly, it imposes minimal data requirements, as it does not necessitate strict adherence to complex distributional assumptions. Consequently, it can achieve satisfactory performance in simpler scenarios. Furthermore, it boasts extensive applicability across diverse contexts.

3. Results and Discussion

3.1 Descriptive Analysis

Fig. 1 presents the heart rate histogram of hypertensive patients. The x-axis of the histogram represents heart rate in beats per minute, while the y-axis denotes frequency. Demonstrate that the heart rate among this population typically averages 80 bpm, which indicates a generally elevated heart rate in individuals with hypertension.

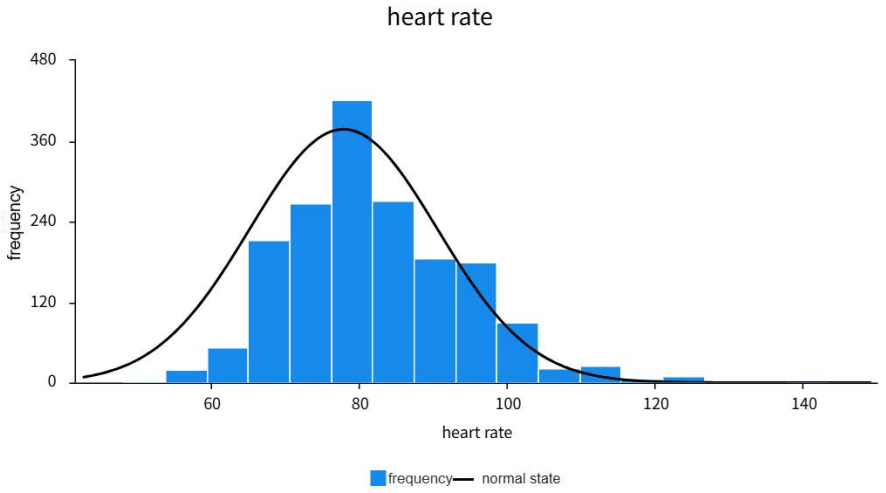

Fig.1 Heart Rate Histogram (Photo/Picture credit: Original).

Fig. 2 presents a histogram illustrating the cholesterol levels of hypertensive patients. The x-axis in the graph represents the cholesterol concentration, measured in milligrams per deciliter (mg/dL), while the y-axis denotes the frequency distribution. The data indicates that the total cholesterol levels in these patients typically reach 250 mg/dL, whereas the normal cholesterol level is approximately 200 mg/dL. This significant disparity suggests a potential correlation between elevated cholesterol levels and hypertension [9].

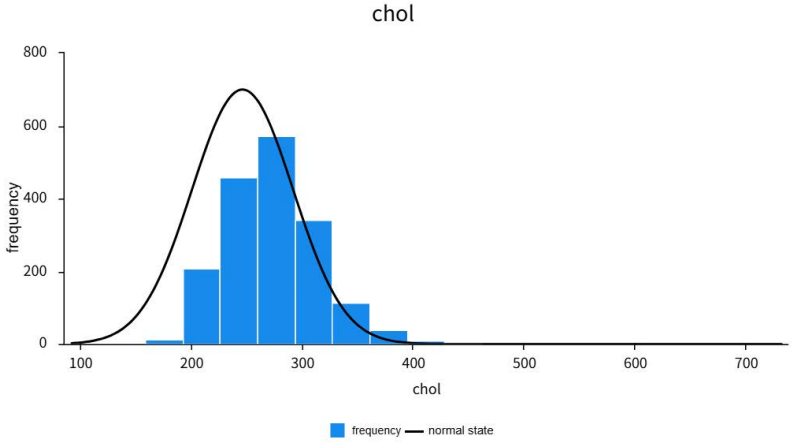

Fig. 2 Chol Histogram (Photo/Picture credit: Original).

Fig. 3 presents the age histogram of hypertension patients. The x-axis in the graph represents the age distribution of individuals within the dataset, while the y-axis denotes the frequency distribution. As illustrated in the histogram, the majority of patients are approximately 50 years old, indicating that the hypertensive population predominantly consists of middle-aged and elderly individuals.

ISSN 2959-6157

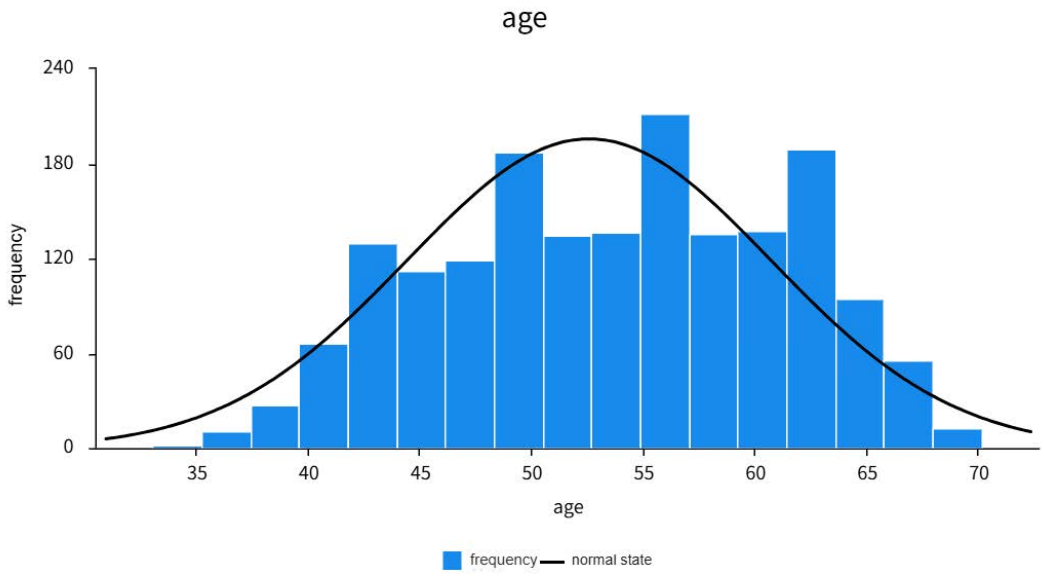

Fig. 3 VAge Histogram (Photo/Picture credit: Original).

Fig. 4 illustrates the BMI histogram of hypertensive patients. The x-axis represents the distribution of individual BMI values within the dataset, while the y-axis denotes the frequency distribution. As depicted in the histogram, the majority of patients exhibit BMI indices exceeding 30, which indicates that this population is classified as obese. This finding suggests a significant correlation between hypertension and obesity within the studied cohort.

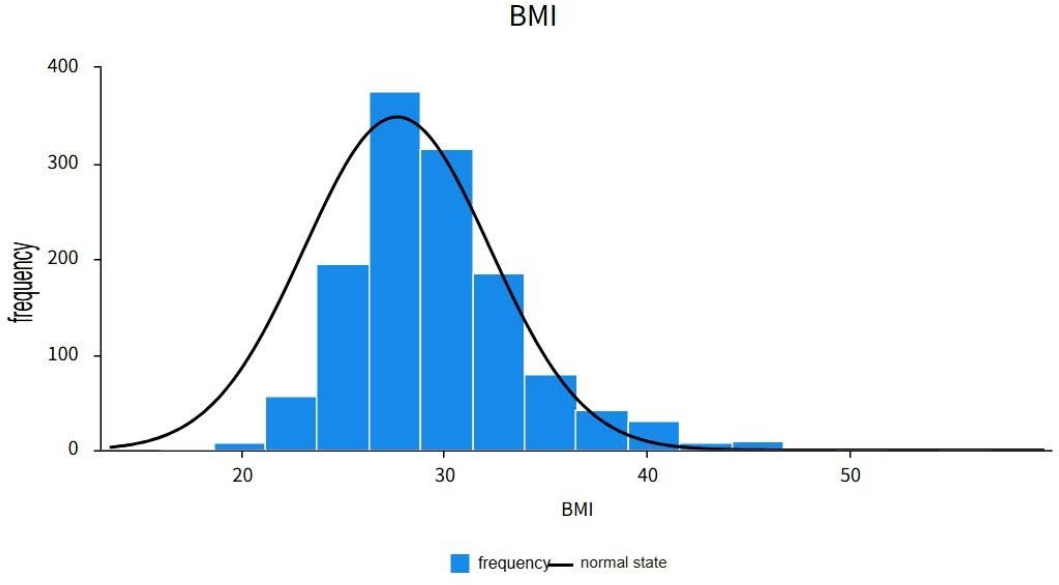

Fig. 4 BMI Histogram (Photo/Picture credit: Original).

3.2 Results of Logistic Regression Analysis

There are numerous influencing factors for hypertension. When studying the impact of smoking, other factors such as age and gender also need to be considered. Simultaneously, elements such as heart rate and cholesterol levels can be utilized to provide a certain degree of predictive assessment for hypertension.

This study considers age, gender, heart rate, cholesterol level, and smoking status as independent variables, with hypertension as the dependent variable. The overall effectiveness of the model was initially analyzed.

As illustrated in Table 2, the null hypothesis of the model test posits that the model quality remains consistent regardless of the inclusion of independent variables (male,

age, currentSmoker, BPMeds, Chol, BMI, heartRate). Given that the p-value is less than 0.05, the null hypothesis is rejected, thereby indicating the efficacy of the incorporated independent variables in the current model construction. This substantiates the meaningfulness of the present model development.

Table 2. Likelihood ratio test results of the binary logit regression model

Model	-2 times the log-likelihood value	Chi-square value	df	p	AIC	BIC
Only intercept	5092.299					
Final model	4060.849	1031.450	7	0.000	4076.849	4127.435

As illustrated in Table 3, a binary logistic regression analysis was conducted with current age, gender, smoking status, heart rate, BMI, cholesterol level, and BPMeds as independent variables, and blood pressure as the dependent variable. The analytical results indicate that male gender, age, BPMeds, cholesterol level, BMI, and heart rate exhibit statistically significant positive correlations with blood pressure, whereas current smoking status demonstrates no significant impact on blood pressure.

Table 3. Summary of binary logit regression analysis results

item	Regression coefficient	Standard error	*"z value"*	Wald χ^2	p value	"OR value"	OR 95% CI
male	0.194	0.081	2.396	5.742	0.017	1.215	1.036 ~ 1.424
age	0.074	0.005	15.092	227.763	0.000	1.077	1.067 ~ 1.087
currentSmoker	-0.077	0.082	-0.929	0.862	0.353	0.926	0.788 ~ 1.089
BPMeds	8.033	3.486	2.305	5.311	0.021	3080.593	3.324 ~ 2855213.389
Chol	0.004	0.001	3.945	15.563	0.000	1.004	1.002 ~ 1.005
BMI	0.158	0.010	15.258	232.794	0.000	1.171	1.148 ~ 1.196
heartRate	0.030	0.003	9.199	84.615	0.000	1.031	1.024 ~ 1.037
Intercept	-11.996	0.510	-23.513	552.884	0.000	0.000	0.000 ~ 0.000
Note: Dependentvariable=BP							
McFadden $R = 0.203$							

As illustrated in Table 4, the predictive accuracy of the binary logit regression model was aggregated, revealing an overall accuracy rate of approximately 75 percent.

Table 4. Summary of binary logit regression prediction accuracy

0		Predicted value		Prediction accuracy rate	Predict error rate
		1			
True value	0	2611	236	91.71%	8.29%
	1	792	480	37.74%	62.26%
Summary				75.04%	24.96%

4. Conclusion

This study focuses on the analysis of the influencing factors of hypertension. By examining the hypertension dataset sourced from the official Kaggle platform, it employs descriptive statistical methods and establishes a binary logistic regression model to investigate the key determinants of hypertension onset, thereby providing predictive insights into the disease.

Descriptive statistical analysis reveals significant correlations between hypertension and several key variables, including BMI, heart rate, cholesterol levels, and age distribution characteristics. Logistic regression analysis further demonstrates that male gender, age, BPMeds,

ISSN 2959-6157

cholesterol, BMI, and heart rate exhibit significant positive associations with blood pressure, whereas current smoking status shows no statistically significant impact. The overall predictive accuracy of this logistic regression model is approximately 75%, with a notably higher accuracy rate of 92% in identifying non-hypertensive individuals. These findings indicate the model's substantial predictive validity. This study elucidates the associations between physiological indicators (such as age, BMI, heart rate, and cholesterol levels) and behavioral factors (including smoking and medication use) with hypertension incidence, providing valuable references for hypertension risk assessment in both general and specific populations. Furthermore, it offers scientific support for developing targeted prevention strategies, such as obesity control, cholesterol management, heart rate monitoring, and smoking cessation interventions. However, this study is subject to certain limitations, as the predictive accuracy of the model requires further improvement, which may be attributed to the constraints of the data sample. Subsequent research should expand the scope of clinical data collection and further refine the model to enhance the predictive accuracy of hypertension onset, thereby providing more robust support for the prevention and management of hypertension.

5. References

[1] Zhao Xiaoxiao, Lu Xiaohui, Ke Lixin, et al. Analysis of the burden of hypertension in the elderly population worldwide and in China from 1990 to 2021 and Prediction of Future Trends. Xiehe Medical Journal, 1-17 [2025-05-23].

[2] Sun, Q., Tian, W., Luo, T., et al. Interpretation of the 2023 World Health Organization Global Hypertension Report. Chinese Journal of Clinical Thoracic and Cardiovascular Surgery, 2024, 31(02): 203-208.

[3] Zhang, D., Li, Y. Summary and Interpretation of the World Health Organization Global Hypertension Report (2023). Journal of Diagnostics Concepts & Practice, 2024, 23(03): 297-304.

[4] Hu, H., Qiu, L. Discussion on the Incidence and Influencing Factors of Hypertension and Health Guidance. Journal of Industrial Medicine, 2012, 25(05): 17-18.

[5] Zhang, Z., Zhao, B., Li, Q., et al. Research Progress on Influencing Factors and Prevention of Hypertension. Health Examination and Management, 2024, 5(02): 157-162.

[6] Liu, F., Ren, M. Research Progress on Smoking-Induced Cardiovascular Diseases. Journal of Xinxiang Medical University, 2023, 40(05): 497-500.

[7] Publisher: Kaggle, Hypertension-risk-model-main Exploring Predictive Factors for Hypertension Risk Prediction, 2024.1.1, 2025.7.1 https://www.kaggle.com/datasets/khan1803115/hypertension-risk-model-main/data

[8] Yu, C., Wang, Y., Chi, X., et al. Similarities and Differences in the Definition, Classification, and Stratification of Hypertension between the 2018 European Society of Hypertension/European Society of Cardiology Guidelines and the Chinese Guidelines for the Prevention and Treatment of Hypertension. Chinese Journal of Hypertension, 2019, 27(09): 811-813.

[9] Bai Guoyan, Peng Lijing, Xie Xuejian, et al. The relationship between lipoprotein cholesterol content and vascular plaque types in patients with cardiovascular diseases of different ages. Modern biomedical progress, 2024, 24 (20): 3947-3949.