Proceedings of CONF-MPCS 2026 Symposium: Theoretic Physics and Plasma Physics
DOI: 10.54254/2753-8818/2026.DL34622

Empirical Analysis of Comorbidity Association Based on Polynomial Logistic Regression

Mingyang Shen

School of Management and E-Business, Zhejiang Gongshang University, Hangzhou, China
smy3236894401@outlook.com

Abstract. Stroke, heart disease, and liver cirrhosis are high-burden chronic diseases in the world, and their comorbidity poses a serious challenge to public health. This paper aims to explore the internal relationship among the three kinds of diseases and provide a scientific basis for clinical comorbidity screening. The relevant open source data sets were selected for the study. After preprocessing, factor analysis and polynomial logistic regression were used to construct the comorbidity risk prediction model. Factor analysis extracted three common factors, cardiovascular risk, metabolism, and gender, with a cumulative variance contribution rate of 82.7%. The results showed that stroke was most significantly associated with liver cirrhosis, with an average joint probability of 7.90% and close to 95% in some individuals. The average probability of heart disease and liver cirrhosis was 4.86%. The probability of three diseases occurring at the same time is low, only 0.49%. The overall accuracy of model prediction was 83.41%, and the classification accuracy of the disease-free and heart disease only combination was more than 90%. The model performance evaluation index shows that its fitting effect is good, so the model has good recognition ability for common disease combinations, and can effectively mine the internal law of multiple disease coexistence.

Keywords: Comorbidity association, Polynomial logistic regression, Factor analysis

1. Introduction

Chronic non-communicable diseases have become a core challenge in the field of global public health. According to the statistics of the World Health Organization, cardiovascular disease, cancer, and other chronic diseases cause nearly 41million deaths every year, accounting for 74% of the total number of deaths in the world [1]. In many chronic diseases, the phenomenon of comorbidity, that is, the same person suffering from two or more chronic diseases at the same time, is increasingly concerned. However, most of the existing studies focus on the risk factor identification and prediction model construction of a single disease, and the discussion on the internal correlation and comorbidity mechanism between different diseases is still relatively scarce. In fact, the formation of comorbidity often involves the interaction of multiple factors, which limits the accurate identification and early intervention of people with multiple diseases to a certain extent.

In the field of disease risk prediction, machine learning methods have shown broad application prospects. Uddin and others compared the performance of a variety of supervised learning

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.DL34622

algorithms in disease prediction and found that the support vector machine was the most frequently used algorithm, while the random forest algorithm showed the highest prediction accuracy, ranking first in 53% of the studies [2]. Mohan et al. used the method of mixed random forest and linear model to achieve an accuracy of 88.7% in the prediction of heart disease [3]. In view of the comorbidity association between hypertension and hyperlipidemia, Di and Xie comprehensively use multiple data mining methods, such as logistic regression and classification and regression tree, to identify common risk factors, and use multiple adaptive regression splines to establish a prediction model, which is far better than the traditional BP neural network [4].

In recent years, researchers have begun to pay attention to the complex association between diseases. Hanlon et al., based on the prospective data analysis of 493737 participants of the British Biological Bank, found that weakness was significantly correlated with the coexistence of multiple diseases. Individuals with four or more long-term diseases had a 27 times higher risk of weakness than those without chronic diseases, and the impact of weakness on mortality was independent of the number of comorbidities [5]. Li and Wang used potential profile analysis to explore the comorbidity mode of depression and anxiety among college students, and found that the comorbidity degree was significantly correlated with social network addiction [6]. Zhang et al. used potential category analysis to classify the preoperative dysfunction of elderly patients with early lung cancer, effectively identified three groups of people: low lung function - low activity tolerance group, anxiety - poor sleep group, and physical weakness group, and analyzed the differences in demographic characteristics of different categories [7]. However, most of the above studies focus on the comorbidity patterns of two types of diseases or specific populations, and the systematic comorbidity association analysis for multiple types of diseases is still relatively limited.

Based on this, this study intends to integrate the multi-source data of stroke, heart disease, and liver cirrhosis, and comprehensively use the methods of factor analysis and polynomial logistic regression to construct the risk prediction model of comorbidity, and explore the internal relationship among the three types of diseases, to provide a scientific basis for clinical comorbidity screening and personalized intervention.

2. Method

2.1. Data sources

The data of this study are selected from the open source data set of chronic diseases provided by a mathematical modeling competition, which includes the stroke data set, heart disease data set, and liver cirrhosis data set [8]. The stroke data set contains 5110 samples, covering 12 characteristics such as age, gender, history of hypertension, blood glucose level, etc. The heart disease data set contains 918 samples, covering 11 characteristics and ECG indicators such as age, chest pain type, resting blood pressure, cholesterol, etc. The liver cirrhosis data set contains 418 samples, including 17 characteristics such as follow-up days, age, bilirubin, albumin, platelets, and disease stage information. Each data set contains the basic information of patients, physiological indicators, and disease diagnosis results, which provides a data basis for the following comorbidity association studies.

2.2. Data preprocessing

For the missing values of numerical characteristics in the data set, in order to adapt to different data distributions, this study uses the mean or median to fill in. For classified features, mode is used to

DOI: 10.54254/2753-8818/2026.DL34622

fill in order to retain the main category information. Then, the features with a missing proportion of more than 30% were deleted. In terms of outlier detection, this study used boxplots and the Z-score method to identify outliers, and combined with the reasonable range of clinical medicine for correction or tail reduction processing. The retention rate of each data set after cleaning was more than 80%. Taking the cholesterol characteristics in the heart disease data set as an example, Figure 1 (a) shows the distribution of the original data, showing a large number of zero value anomalies; Figure 1 (b) shows the distribution after eliminating zero values, and the data is more in line with clinical practice. Finally, in order to eliminate the dimensional effect, all numerical features are standardized. After the above pretreatment, 4908, 746, and 399 valid samples were retained in the data sets of stroke, heart disease, and liver cirrhosis, respectively.

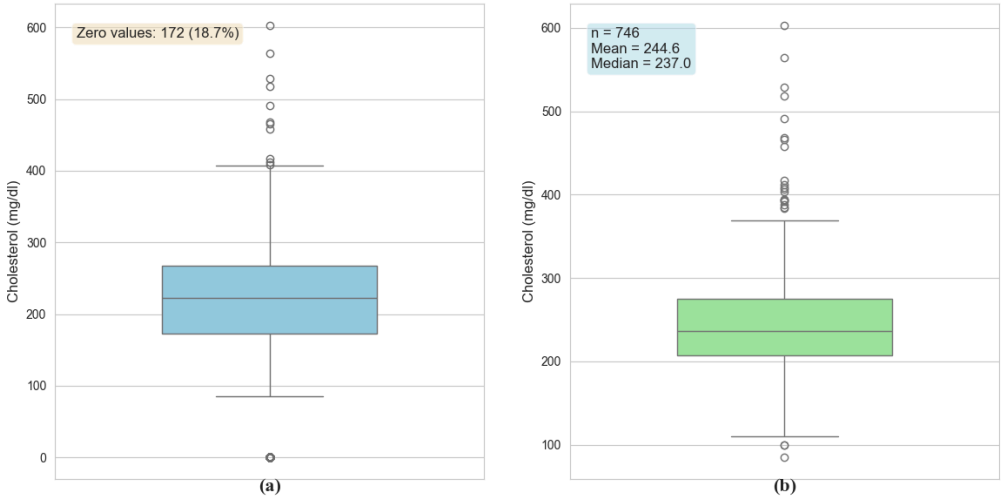

Figure 1. Cholesterol distribution boxplot: (a) Before correction. (b) After correction (photo/picture credit: original)

2.3. Data processing method

In this study, an integrated analysis framework of factor dimensionality reduction, multi-classification modeling, and joint probability derivation was constructed to explore the law of chronic disease comorbidity. First, this study screened common risk characteristics from multi-source data and verified the applicability of factor analysis through the KMO sampling adequacy measure and Bartlett's sphericity test. The higher the KMO value, the stronger the partial correlation between variables. The Bartlett test is significant; the variables are not independent. The combination of the two can objectively judge whether the data is suitable for dimensionality reduction [9, 10]. Subsequently, factor analysis is used to extract common factors, reveal the potential risk structure while eliminating the feature multicollinearity, and take the factor score after dimension reduction as the model input [11]. On this basis, this study builds a polynomial logistic regression model, which predicts the probability of multiple disease combinations at the same time through the softmax function, can improve the prediction accuracy by using the correlation between outputs, and introduces L2 regularization to control overfitting, and finally obtains the joint probability of four comorbidity combinations, thus providing a complete implementation path for the study of comorbidity association [12].

DOI: 10.54254/2753-8818/2026.DL34622

3. Result

3.1. KMO and Bartlett's sphericity test

As shown in Table 1, the overall KMO value is 0.75, indicating that there is a strong partial correlation between variables, which is suitable for factor analysis. The approximate chi-square value of Bartlett's sphericity test was 287.4 (p-value < 0.001), which further verified the feasibility of factor analysis.

Table 1. KMO and Bartlett's sphericity test results

KMO value		0.751
Bartlett's sphericity test	Approximate chi-square	287.4
	Degrees of freedom	10
	Significance	0.000

Figure 2 shows the MSA value of each variable. The abscissa represents five risk variables, such as age, and the ordinate represents the MSA value. Its value range is 0-1, which is used to measure the common degree of variation between each variable and other variables. The MSA values of all variables in Figure 2 are higher than 0.7. Among them, the MSA value of hypertension is the highest, indicating that there is a strong common variation between this variable and other characteristics. Age and blood glucose were the second, indicating that they were important potential factors affecting the coexistence of multiple diseases. Although BMI and gender were slightly lower than other variables, they were still significantly higher than 0.7, indicating that they also had stable contributions to public factors. In summary, all variables are suitable for inclusion in subsequent factor analysis without removing any features.

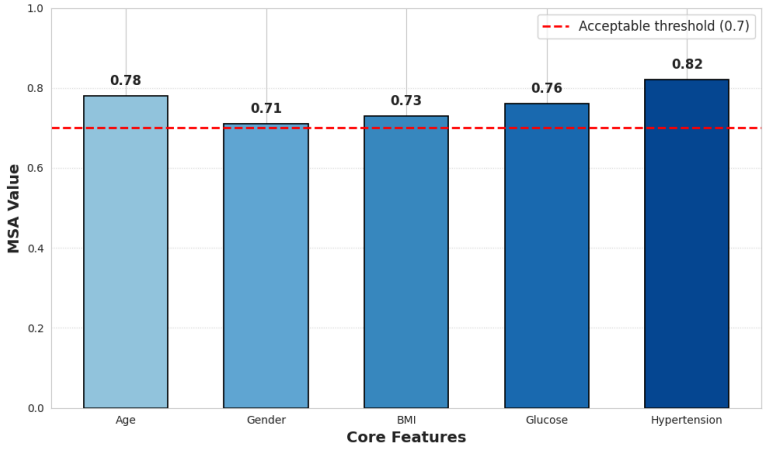

Figure 2. MSA value of each variable (photo/picture credit: original)

3.2. Factor analysis

Let the original characteristic matrix be $X \in \mathbb{R}^{n \times p}$, and the factor analysis model can be expressed as:

$$X = \mu + F\Lambda^T + \varepsilon \tag{1}$$

DOI: 10.54254/2753-8818/2026.DL34622

Where $F \in \mathbb{R}^{n \times k}$ is the factor score matrix, $\Lambda \in \mathbb{R}^{\mathrm{p} \times \mathrm{k}}$ is the factor load matrix, k is the number of common factors, and ε is the error term. In this study, three common factors were extracted by retaining the factors with eigenvalues greater than 1, and the cumulative variance contribution rate was 82.7%. As shown in Table 2, factor 1 has a high load on hypertension and age, which can be explained as a cardiovascular risk factor. Factor 2 has a high load on BMI and blood glucose, which reflects the metabolic state and can be expressed as a metabolic factor. Factor 3 is strongly correlated with gender, which is called the gender factor.

Table 2. Factor load matrix

Feature	Factor 1	Factor 2	Factor 3
Age	0.745	0.213	-0.102
Gender	-0.123	0.089	-0.921
BMI	0.323	0.531	0.187
Glucose	0.297	0.535	0.065
Hypertension	0.689	0.241	-0.118

3.3. Polynomial logistic regression model

In this study, three common factors were used as inputs, and eight disease combinations were used as outputs. For the i -th sample, if its factor score vector is $F_i \in \mathbb{R}^3$, the probability that it belongs to the j -th disease combination is given by the softmax function. Finally, the joint state probability function is obtained as follows:

$$P(Y_i = j|F_i) = \frac{exp\left(\beta_{j0}+\beta_j^T F_i\right)}{\sum_{m=1}^{8} exp(\beta_{m0}+\beta_m^T F_i)} \tag{2}$$

Where Y_i is the disease combination label of sample i , $j \in \{0,1,\ldots,7\}$, β_{j0} is the intercept term of the j -th combination, β_j^T is the weight coefficient vector transposition of the j -th combination, β_{m0} and β_m^T are the corresponding parameters of the m -th combination.

As shown in Table 3, the model has the strongest ability to identify the disease combination of label 0 and label 2. The accuracy rate of the comorbidity category is slightly low due to the small sample size, but the overall misjudgment rate is less than 20%, meeting the reliability requirements of clinical risk assessment.

Table 3. Classification accuracy of each disease combination

Disease combination	Label	Classification accuracy
Disease free	0	92.3%
Stroke	1	78.6%
Heart disease	2	90.1%
Liver cirrhosis	3	81.4%
Stroke + Heart disease	4	82.5%
Stroke + Liver cirrhosis	5	79.8%

DOI: 10.54254/2753-8818/2026.DL34622

Table 3. (continued)

Heart disease + Liver cirrhosis	6	80.2%
All diseases	7	71.0%

3.4. Joint probability prediction of comorbidity

In order to obtain the joint probability of any two or three diseases occurring at the same time, this study sums the probability of the corresponding mutually exclusive categories. Let S, H and C represent stroke, heart disease, and liver cirrhosis, respectively, and $Y\ represents$ the disease combination label, then the joint probability calculation formula is as follows:

$$\begin{cases} P(S \cap H) = P(Y=4) + P(Y=7) \\ P(S \cap C) = P(Y=5) + P(Y=7) \\ P(H \cap C) = P(Y=6) + P(Y=7) \\ P(S \cap H \cap C) = P(Y=7) \end{cases} \tag{3}$$

Based on a total of 1289 samples in the test set, the joint probability of each disease combination was calculated and obtained. As shown in Table 4, the risk of comorbidity of stroke and liver cirrhosis is the highest, with an average probability of 7.90%, and the risk for some individuals is close to 95%, indicating that the risk difference between individuals in this combination is significant. The average probability of comorbidity of heart disease and liver cirrhosis was 4.86%. The risk of stroke and heart disease was 3.85%. The probability of three diseases occurring at the same time is very low, only 0.49%.

Table 4. Descriptive statistics of comorbidity joint probability

Comorbidity type	Comorbidity type	Maximum probability	Standard deviation
Stroke + Heart disease	3.85%	54.48%	0.15
Stroke + Liver cirrhosis	7.90%	94.63%	0.25
Heart disease + Liver cirrhosis	4.86%	49.54%	0.18
All diseases	0.49%	4.95%	0.05

3.5. Model performance evaluation

In order to further quantify the prediction accuracy of the model for the joint probability of comorbidity, five indicators, such as MSE and RMSE, were used to evaluate the prediction results of the test set. Among them, MSE and RMSE are sensitive to large errors and can reflect the overall deviation of the predicted value; MAE directly measures the magnitude of average absolute error; MAPE presents relative error in the form of a percentage, which is convenient for cross-model comparison; R^2 represents the degree to which the model explains the real variability. As shown in Table 5, all evaluation indices show that the model has a good fitting effect and prediction reliability, and the error is controlled within the acceptable range. R^2 is close to 0.7, indicating that the model can explain about 70% of the data variation and has good generalization ability.

DOI: 10.54254/2753-8818/2026.DL34622

Table 5. Model regression evaluation index

Index	MSE	RMSE	MAE	MAPE	R^2
Value	0.021	0.145	0.108	12.36%	0.683

4. Conclusion

For stroke, heart disease, and liver cirrhosis, this study constructed a comorbidity prediction framework combining factor analysis and polynomial logistic regression. Factor analysis successfully extracted three core risk dimensions, with a cumulative variance contribution rate of 82.7%, revealing the key role of metabolic abnormalities in the pathogenesis of multiple diseases. The overall prediction accuracy of the model is 83.41%, and the recognition performance of common single or double disease combinations is particularly robust.

Joint probability analysis further showed that there was a significant non-random comorbidity association among the three types of chronic diseases. Among them, the risk of comorbidity of stroke and liver cirrhosis is the most prominent, followed by heart disease and liver cirrhosis, and the probability of three diseases occurring at the same time is very low, less than 0.5%. This finding suggests that the association between stroke and cirrhosis may be mediated by common metabolic and vascular pathological mechanisms. In clinical practice, it should pay attention to the evaluation of liver function in patients with ischemic cerebrovascular disease, and vice versa. To sum up, the analysis framework proposed in this study can effectively explore the internal law of multiple disease coexistence and provide a quantitative basis for differential comorbidity screening and personalized intervention.

Although this study effectively revealed the comorbidity association rules among the three types of diseases, there are still some limitations, such as unbalanced sample size and limited variable coverage. Future research can further introduce large sample data from the real world, and combine more abundant characteristic dimensions, such as lifestyle and genetic information, to improve the prediction performance and clinical applicability of the model. At the same time, exploring the application potential of deep learning methods in comorbidity pattern recognition is helpful to capture the more complex nonlinear relationships in the data. On this basis, the development of an interactive risk assessment tool will provide more intuitive and convenient decision support for clinical practice.

References

[1] World Health Organization. (2023, May 25). World health statistics 2023. https: //ewitkey.cn/cms/show-31282.html
[2] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1), Article 281.
[3] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554.
[4] Di, X. M., & Xie, H. W. (2013). On mining common risk factors of multi-diseases and MARS prediction model. Computer Applications and Software, 30(10), 36–40.
[5] Hanlon, P., Nicholl, B. I., Jani, B. D., McQueenie, R., Gallacher, K. I., Cullen, A. M., ... & Mair, F. S. (2018). Frailty and pre-frailty in middle-aged and older adults and its association with multimorbidity and mortality: A prospective analysis of 493 737 UK Biobank participants. The Lancet Public Health, 3(7), e323–e332.
[6] Li, J., & Wang, S. P. (2026). Latent profile analysis of comorbidity of depression and anxiety symptoms in college students and its correlation with social networking service addiction. Journal of Shanghai Jiao Tong University (Medical Science), 1–7.

DOI: 10.54254/2753-8818/2026.DL34622

[7] Zhang, R., Yang, Z., Shen, X., & Li, Y. (2024). Preoperative physical dysfunction characteristics and influence factors among elderly patients with early lung cancer: A latent class analysis. Medical Devices: Evidence and Research, 17, 1743–1754.
[8] Saikr. (2025, July 11). 2025 Asia and Pacific Mathematical Contest in Modeling. https: //new.saikr.com/vse/apmcm/202502
[9] Franić, T. (2024). Translation, cross-cultural adaptation, and validation of the Croatian version of the Athlete Psychological Strain Questionnaire (APSQ). Sports, 12(3), Article 61.
[10] Sun, S., Liu, H., Guo, Z., ... & EPMA Short-form Suboptimal Health Status Questionnaire Validation Study (ESSV) Group. (2023). Development and validation of a short-form suboptimal health status questionnaire. The EPMA Journal, 14(4), 601–612.
[11] Lian, J., Kiely, K. M., & Anstey, K. J. (2022). Cumulative risk, factor analysis, and latent class analysis of childhood adversity data in a nationally representative sample. Child Abuse & Neglect, 125, Article 105486.
[12] Pate, A., Riley, R. D., Collins, G. S., van Smeden, M., Van Calster, B., Ensor, J., & Martin, G. P. (2023). Minimum sample size for developing a multivariable prediction model using multinomial logistic regression. Statistical Methods in Medical Research, 32(3), 555–571.