Diagnosis of Hepatobiliary Disease Based on Logistic Regression Model

Xin Gai\textsuperscript{1,}\textsuperscript{*}, Yimin Zhang\textsuperscript{2}

\textsuperscript{1}Zhejiang University City College, Zhejiang, China
\textsuperscript{2}Zhejiang University City College, Zhejiang, China

*Corresponding author e-mail: 849733413@qq.com

Abstract: Based on the monitoring data of patients with hepatobiliary disease, we classified and selected four biomarkers: total bilirubin, albumin, alkaline phosphatase, and alanine aminotransferase. A quaternary logistic regression model was established and the model was tested for likelihood ratio and accuracy. Based on the test results, we further optimized the model and established a ternary ungrouped logistic model, and conducted a series of tests of saliency and accuracy. Secondly, according to the data collected from a hospital's real medical examination, the ternary logistic model was used to give predictions for the prevalence. Because the ternary logistic model factors were selected and fixed, the sickness Status were integrated, and a binary ungrouped logistic regression model was established. Based on the ROC curve evaluation method, the performance of the model was verified: The model is highly reliable for negative results and has certain reliability for positive results. This large savings in monetary costs, to some extent reduce the risk of disease. Finally, we qualitatively analyze the two established models and draw their own advantages and disadvantages: The prediction effect of this ternary logistic regression model is great, besides, it has satisfactory accuracy and strong robustness, but the elasticity of the selected factors is small; the binary ungrouped Logistic regression model has a good prognostic effect, the elasticity of selection factors is large and the scope of application is wide. Both models have good performance in the preliminary judgment of hepatobiliary diseases.

1. Research Background
As of December 31, 2016, there were more than 433 million major cases of chronic hepatobiliary disease in China. The overall incidence of major hepatobiliary diseases in China increased at a CAGR of 0.5% from 2012 to 2016 and is expected to increase at a CAGR of 0.6% from 2016 to 2021.

The prevalence of hepatobiliary growth is mainly caused by poor living habits. Smoking is a risk factor for the development of hepatocellular carcinoma and can increase the risk of liver cancer in patients with hepatitis B. In addition to its own harm, smoking can also "hand in hand" to drink, accelerate the deterioration of the patient's disease, and ultimately lead to the tragedy. However, early treatment of hepatobiliary disease is relatively easy. As long as active treatment and improvement of living habits are detected in the early stage, the cure rate of hepatobiliary disease is over 90%.

In recent years, as the concept of healthy life has gained deep roots in the hearts of people, the regular medical examination rate has risen. However, the variety of physical examinations varies, and the inspection items provided are also quite different. Based on this background, we team will thoroughly explore the preliminary diagnostic methods for hepatobiliary diseases under various test
items (factors) in order to achieve a high accuracy of the identification of hepatobiliary diseases and improve the hope of healing of patients with hepatobiliary diseases.

2. Research methods and model solving

2.1. Quaternary ungrouped logistic regression model

By observing the data of patients with hepatobiliary disease and looking through medical knowledge, we can know that the disease types of the sample population can be divided into: liver disease, chronic gall disease, gall disease, and illness. So consider creating a multinomial logistic regression model:

\[ P(y = k | x) = \frac{e^{\beta_k(x)}}{1 + \sum_{j=0}^{c-1} e^{\beta_j(x)}} \]  \hspace{1cm} (1)

Among them \( k = 0,1,2,\cdots,c-1 \). The corresponding logistic regression model can be derived from this:

\[ g_k(x) = \ln \left( \frac{P(y = k | x)}{P(y = 0 | x)} \right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p \]  \hspace{1cm} (2)

The data set variable declaration is as follows:

| Model | Variable Description |
|-------|----------------------|
| X1    | Total bilirubin(T-BIL) |
| X2    | Albumin(ALB) |
| X3    | Alkaline phosphatase(ALP) |
| X4    | Alanine transaminase(ALT) |
| Y     | liver disease | chronic gall disease | gall disease | illness |

The kappa () function is used to test the multiple linearity of the data: k = 5.7933591, and there is no multilinearity among the independent variables. [1]

Through R (multinom function) [2], regression modeling of the sample data set (Appendix 2), the following results are obtained:

| model | intercept | \( \hat{\beta}_1 \) | \( \hat{\beta}_2 \) | \( \hat{\beta}_3 \) | \( \hat{\beta}_4 \) |
|-------|-----------|-----------------|-----------------|-----------------|-----------------|
| y=2   | -9.4379   | -0.0115         | 0.2799          | 0.0005          | -0.0201         |
| y=3   | -6.0355   | -0.0143         | 0.1655          | 0.0067          | -0.0134         |
| y=4   | -33.4323  | -0.1103         | 0.9200          | -0.0472         | -0.0136         |

The p-value of the likelihood ratio test is obtained by using R as follows:
Table 3. Quaternary logistic regression model coefficients test p-value.

| coefficient | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ |
|-------------|----------------|----------------|----------------|----------------|
| Test P value | 2.98E-05 | 0 | 3.26E-06 | 3.98E-09 |

Factor coefficients pass the significance test. Next, we use the predict function to find the fitted values, and plot the pie charts in the four cases with the actual values as follows:

**Figure 1.** Fit Result Pie Chart in Four Different Cases.

Synthesize four kinds of situations:

**Figure 2.** Quaternary comprehensive fitting result pie chart.

According to the above figure, it can be seen that the fitting effect on gall disease (anxious and slow gall) is not good. When we look up the data, we find that there is a high correlation between gall diseases, so we no longer need to be able to identify the gall disease clearly, while the classification of the dependent variable was changed (liver disease, gall disease, normal). Therefore, we next set up a logistic regression model with Ternary ungrouped logistic model.

2.2. Ternary ungrouped logistic model

The classification of the dependent variable was changed to (liver disease, gall disease, and normal), and a logistic regression model with ternary ungrouped logistic model was established. In order to facilitate the fitting of the model, we randomly stratified the data of patients with hepatobiliary disease and divided the original data set into training data sets and test data sets. The data set variable declaration is as follows:
Table 4. Model II Data Description

| variate | meaning                                    |
|---------|--------------------------------------------|
| X1      | Total bilirubin(T-BIL)                     |
| X2      | Albumin(ALB)                               |
| X3      | Alkaline phosphatase(ALP)                  |
| X4      | Alanine transaminase(ALT)                  |
| Y       | 1  liver disease                           |
|         | 2  gall disease                            |
|         | 0  normal                                  |

Make:

\[ \pi_i = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)} \]  

Through R, regression modeling of the sample data set (Appendix 2) is performed first, and the following results are obtained:

Table 5. Ternary ungrouped logistic model coefficients.

| model | intercept | \( \hat{\beta}_1 \) | \( \hat{\beta}_2 \) | \( \hat{\beta}_3 \) | \( \hat{\beta}_4 \) |
|-------|-----------|----------------------|----------------------|----------------------|----------------------|
| y=1   | 31.0623   | 0.1067               | -0.8525              | 0.0397               | 0.0150               |
| y=2   | 23.4051   | 0.0943               | -0.6198              | 0.0461               | -0.0025              |

Table 6. Ternary ungrouped logistic model standard error.

| model | intercept | \( \hat{\beta}_1 \) | \( \hat{\beta}_2 \) | \( \hat{\beta}_3 \) | \( \hat{\beta}_4 \) |
|-------|-----------|----------------------|----------------------|----------------------|----------------------|
| y=1   | 4.6250    | 0.0281               | 0.1098               | 0.0125               | 0.0115               |
| y=2   | 4.2681    | 0.0273               | 0.1002               | 0.0119               | 0.0105               |

And the residual error value is 259.0633 and the AIC value is 279.0633.

The test significance is obtained by:

Table 7. Ternary ungrouped logistic model coefficient test p-value.

| coefficient | \( \hat{\beta}_1 \) | \( \hat{\beta}_2 \) | \( \hat{\beta}_3 \) | \( \hat{\beta}_4 \) |
|-------------|----------------------|----------------------|----------------------|----------------------|
| Test P value| 7.18E-05             | 0                    | 1.93E-05             | 1.62E-08             |

The factor coefficient passes the significance test and the equation is obtained from the above result:

For \( y=1 \) (liver disease)

\[ \pi_1 = \frac{\exp(31.0623 + 0.1067 x_1 - 0.8525 x_2 + 0.0397 x_3 + 0.015 x_4)}{1 + \exp(31.0623 + 0.1067 x_1 - 0.8525 x_2 + 0.0397 x_3 + 0.015 x_4)} \]  

For \( y=2 \) (gall disease)

\[ \pi_2 = \frac{\exp(23.4051 + 0.0943 x_1 - 0.6198 x_2 + 0.0461 x_3 - 0.0025 x_4)}{1 + \exp(23.4051 + 0.0943 x_1 - 0.6198 x_2 + 0.0461 x_3 - 0.0025 x_4)} \]  

Next, the test data set is predicted, and the pie charts are plotted in the following three cases in combination with the actual values:
Synthesize three situations of fitting results:

**Figure 3.** Fitting Results Pie Charts in Three Different Cases.

From the above figure, we can directly see that the accuracy of the established ternary ungrouped logistic model is very high. As long as the physical examination data of total bilirubin (T-BIL), albumin (Alb), alkaline phosphatase (ALP), and alanine aminotransferase (ALT) are obtained from a certain population, it can be established well whether or not he suffers from hepatobiliary diseases to conduct the preliminary screening.

Now, we have a group of comprehensive physical examination data from a certain hospital, including: gender, age, systolic blood pressure, diastolic blood pressure, pulse, weight, height, past history, smoking, alcohol consumption, surgical history, hepatitis B surface antigen (HBsAg), total cholesterol (TC), triglycerides (TG), alanine aminotransferase (ALT), aspartate aminotransferase (AST), total bilirubin (T-BIL), indirect bilirubin (IB), alkaline phosphatase (ALP), test items such as total protein (TP), albumin (Alb), and globulin (GLB). Below, we use this model to predict the prevalence of hepatobiliary disease in a physical examination group in a hospital. The results are as shown in Table 8.
Table 8. Outcomes of Hepatobiliary Disease in a Hospital Experience Population.

|    | 0  | 2  | 0  | 0  | 0  | 0  | 2  | 2  | 0  | 2  |
|----|----|----|----|----|----|----|----|----|----|----|
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 0  | 2  | 0  | 0  | 0  | 2  | 0  | 0  | 0  |
|    | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 2  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 0  | 2  | 0  | 0  | 0  | 1  | 0  | 0  | 0  |
|    | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 2  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 2  | 2  | 2  |
|    | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 2  | 0  | 2  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 2  | 0  | 0  | 0  |
|    | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 2  | 0  | 2  | 0  | 0  | 0  | 2  | 0  | 0  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 0  | 0  | 2  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 2  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 2  | 0  | 0  | 0  | 0  | 0  | 0  | 2  | 0  | 2  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 2  |
|    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|    | 0  | 0  | 2  | 0  | 0  | 0  | 2  | 0  | 0  | 2  |

(values of 0, 1, and 2 indicate normal, liver, and bile disease, respectively)

2.3. Binary ungrouped logistic regression model

2.3.1. Model establishment

Since our country began to effectively control epidemic and infectious hepatitis, the incidence of hepatitis has been decreasing year by year. Observing the results, we found that the discrimination result was 1 and that there was only 1 patient with hepatitis, which was consistent with the actual situation.

The general physical examinations in ordinary schools, companies, etc. may not include so many inspection items. We try to establish another model to judge the test of hepatobiliary disease. Taking into account the actual situation of the sample data, it can be regarded as an approximate dependent variable for the two categories of data. Therefore, a Binary ungrouped logistic regression model was established:

$$logit(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = x_i' \beta, \quad i = 1, \ldots, n$$

Hypothesis $y$ for $0-1$ variable, is $y_i \sim B(1, \pi_i)$ and $x_1, \ldots, x_p$ is the certain variables that has an effect on $y$. In $(x_{i1}, \ldots, x_{ip})$ of Different points $(x_{i1}, \ldots, x_{np})$, $i = 1, \ldots, n$, correct $y$ Conducted $n$ Observed independent observations $y_{i1}, \ldots, y_{in}$. Obviously, $y_{i1}, \ldots, y_{in}$ are mutually independent Bernoulli random variables, the probability distribution is:

$$\pi_i y_i (1-\pi_i)^{1-y_i}, \quad y_i = 0,1$$

(7)

Then, the likelihood function of $y_{i1}, \ldots, y_{in}$ is:

$$L(\pi_1, \ldots, \pi_n) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

(8)
Its log-likelihood function is:

\[
l(\pi_1, \cdots, \pi_n) = \sum_{i=1}^{n} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]
\]

(9)

Substitute the above formula into the probability distribution function:

\[
l(\beta) = \sum_{i=1}^{n} [y_i x'_i \beta - \log(1 + \exp(x'_i \beta))]\]

(10)

Begging the maximum likelihood estimate of \( \beta \) is to find \( \beta \) which make \( l(\beta) \) to reach maximum. To do this, calculate the formula about the first derivative of \( \beta \):

\[
\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}}) x_i = X' e
\]

(11)

Among them

\[
X = (x_1, \cdots, x_n)' , e = (e_1, \cdots, e_n)' , e_i = y_i - \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}}
\]

(12)

Make

\[
X' e = \sum_{i=1}^{n} (y_i - \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}}) x_i = 0
\]

(13)

But the above equation about parameters \( \beta \) is a more complex nonlinear function, to get \( \beta \) Maximum Likelihood Estimation \( \hat{\beta} \) is not easy. In general, iterative algorithms, such as the Newton-Raphson iterative algorithm, can be used to find numerical solutions.

The data set variable declaration is as follows:

**Table 9. Model III Data Description**

| variate | meaning                                      |
|---------|----------------------------------------------|
| X1      | Total cholesterol(TC)                        |
| X2      | Triglyceride(TG)                             |
| X3      | Aspartate                                    |
| X4      | aminotransferase(AST)                        |
| X5      | Indirect bilirubin(IB)                       |
| X6      | Globulin(GLB)                                |
| X7      | gender                                       |
| X8      | age                                          |
| X9      | systolic blood pressure                      |
| X10     | pulse                                        |
| X11     | weight                                       |
| X12     | height                                       |
| X13     | smoking                                      |
| Y       | alcohol consumption                          |
| Y       | smoking                                      |
| Y       | alcohol consumption                          |

|       | positive                                      |
| Y       | negative                                     |
The kappa () function is used to test the multiple linearity of the data: $k = 17.47144$, and there is no multilinearity among the independent variables.

We use the glm() function in the R software package DAGG to solve the problem (Appendix 3). After the variable filtering (backward method), the following results are obtained:

Table 10. Binary ungrouped logistic regression model results.

| model | Estimate | Std. Error | z value | Pr(>|z|) |
|-------|----------|------------|---------|----------|
| intercept | 3.109 | 1.97335 | 1.575 | 0.1151 |
| $x_1$ | -0.44004 | 0.22486 | -1.957 | 0.0504 |
| $x_4$ | 0.08137 | 0.04229 | 1.924 | 0.0543 |
| $x_9$ | -0.04710 | 0.02222 | -2.12 | 0.0340 |

And the AIC value is 156.34.

It can be seen from the test results that the P values of $x_1$ and $x_4$ are close to 0.05, and the P value of $x_9$ is less than 0.05. It can be considered that each coefficient of the model is significant, and the equation is obtained:

$$P = \frac{\exp(3.109 - 0.44004x_1 + 0.08137x_4 - 0.0471x_9)}{1 + \exp(3.109 - 0.44004x_1 + 0.08137x_4 - 0.0471x_9)}$$

(14)

2.3.2. Model checking

So how does the equation work? Next we will discuss this issue.

One type of test is utilized to confirm a certain event. It is a qualitative method and is clinically very common. As a rapid and economical screening method, the results are expressed as positive and negative, and the existence of the event is identified. Positive is a positive result of the detection of a disease, meaning that it is determined to be sick. The negative result is determined to be not sick. [3]

However, when the test is inaccurate, one positive may be a false positive and the negative result may also be a false negative. A method yields a statistical number of results after measuring multiple individuals. There are four ways to verify and confirm the results of the method of using other methods:

- **True positive**: the result is accurate and really sick;
- **False positive**: the result is wrong, and it is actually not sick;
- **True negative**: the result is accurate, really no disease;
- **False negative**: the result is wrong, actually sick.

The "sensitivity" and "reliability" of such test results is to evaluate the usefulness of this screening method.

- Sensitivity is the proportion of correctly positive = true positive / (true positive + false negative) = probability that the positive result of the sample is correct.
- Specificity is the proportion of true negatives = true negatives / (true negatives + false positives) = the probability that the sample negative conclusion is correct.

We implement this test method based on the R(pROC package) (Appendix 4) to obtain:
According to the above figure, the proportion of correct positives is 0.5, the proportion of correct negatives is 0.892, and the probability threshold is 0.201. It can be seen that the model is highly reliable for negative results and has certain reliability for positive results. Based on our background, after a preliminary screening result for hepatobiliary disease is negative, it is generally not tested in depth. According to model reliability, the risk of illness is indeed very small, and there is no need for further testing or economic costs. When the initial screening results for hepatobiliary disease are positive, people often check further, and the correct rate of positive is 50%, which expands the range of risk population and reduces the risk of disease. Moreover, the proportion of 50% is relatively large and does not cause greater waste.

Let's look at the prediction effect of the model outside the sample (Appendix 5), and plot the pie chart under the three conditions in combination with the actual value as follows:

\[ \text{Fit result of } y = 1 \text{ (positive, hepatobiliary disease)} \quad \text{Fit result of } y = 0 \text{ (negative, illness)} \]

**Figure 6.** Fit Results Pie Chart in Two Different Cases.

Integrate the fitting effect of two situations:

**Figure 7.** Binary comprehensive fitting result pie chart.
From the above results, it can be seen that the prediction effect on the data outside the sample is better, and the screening rate for patients without hepatobiliary disease (negative) is higher. The reason why the model does not achieve the expected results is because there are insufficient data outside the sample, which may cause uneven distribution of samples and high contingency.

So we believe that the model can also be used as a reliable hepatobiliary disease judgment model.

3. Conclusion

Based on the two hepatobiliary disease prediction models we established in the previous section, we find that the ternary ungrouped logistic model has good prediction accuracy, satisfactory accuracy and strong robustness, but it has little flexibility in the selection of factors. When there were four items of total bilirubin, albumin (Alb), alkaline phosphatase (ALP), and alanine aminotransferase (ALT) in a physical examination, we selected this model for preliminary screening of hepatobiliary diseases.

The established binary ungrouped logistic regression model has better predictive effect and more flexible selection of factors. When the physical examination items do not meet the four factors of the ternary ungrouped logistic model, other variables can be selected to use the binary ungrouped logistic regression model to diagnose hepatobiliary diseases.

At present, our country is still a country with a high incidence of hepatobiliary disease. However, if the unusual hepatobiliary disease is discovered in time, it is often prone to worsen or even cancer. Through our research, it has been found that hepatobiliary diseases can be judged to a large extent by the physical examination. Therefore, it is hoped that we can cherish the body, perform regular medical examinations, and annihilate the lesions in the bud.

Thanks Hongpeng Yuan, Yueqin Wei, Qiangwei Ge for helping with this work.

References

[1] Yi Xue, Liping Chen, R statistical modeling and R software, Tsinghua University Press, Beijing, 2007, pp.304-307
[2] Xiaoqun He, Application Regression Analysis (R Language Edition), Publishing House of Electronics Industry, Beijing, 2017, pp.241-242
[3] Information on http://blog.sina.com.cn/s/blog_70978d1e0100wtmq.html