A Statistical Analysis of the Likelihood of Heart Disease and the Recommended Medical Tests

Yifan Wu MS Biostatistics & Data Science

BACKGROUND

It is essential to conduct the early diagnosis to prevent the heart related deaths. After conducting the GLM and CLM model with ROC and AUC and check the odds ratio with 95% confidence intervals, we can the make a choice on medical tests. The combination of medical tests should be restecg, exang, oldpeak, and slope tests. The total cost of these medical test is $102.8 ($15.50 for restecg test and $87.30 for exang, oldpeak, and slope tests). This cost is relatively low and acceptable even if the one does not have the insurance. We have to admit that the limitation of this study is that some variables do not meet the PH assumptions. This may affect the results and need the further study.

OBJECTIVES

Cardiovascular diseases are the number 1 cause of death in adults in the United States. It is crucial for people to prevent heart related deaths by early diagnosis. Doctors make a diagnosis of the disease based on multiple medical test that potential cardiovascular patients take. Even though individual doctors have well prepared knowledge and experiences for the accuracy of the diagnosis, there is still need to improve the accuracy of the diagnosis by constructing a fitting model that takes the results from various medical tests into consideration. The objective of this study is to conduct appropriate hypothesis tests, develop statistical models and do a statistical analysis of the likelihood of heart disease and the recommended medical tests. Typically, for the heart disease outcome measure, it is required to build a model with the severity of heart disease as response variable. Moreover, the outcome of the severity of the heart disease need dichotomized based on the presence of narrowing vessel.

METHODS

Two US hospitals and two European hospitals provide the total 920 patients information for the study. Both demographic and diagnostic information characterize the different conditions of patients. Personal information are based on two variables, age, sex and area. Diagnosis related information is stated in the rest of variables, including cp, trestbps, fbs, restecg, thalach, exang, oldpeak, slope, ca, and thal. The outcome of interest in this study is the variable diag. The variable diag is coded corresponding to the low to high severity from 1 to 4 and the condition of no heart disease. In the later study, one variable, dichotomized_diag, is introduced to identify if a patient has the heart disease or not.

There is no missing value in the data set. In the descriptive analysis, contingency tables are constructed for both continuous and categorial variables. In terms of continuous variables, such as age, trestbps, thalach, oldpeak and ca, some statistics like the mean and variance are first reviewed. A two-sample t test is used to determine the potential relationship between the diagnosis of heart diseases and those continuous variables. For the rest of categorial variables, such as sex, cp, fbs, restecg, exang, slope, thal and area,

METHODS (Continued)

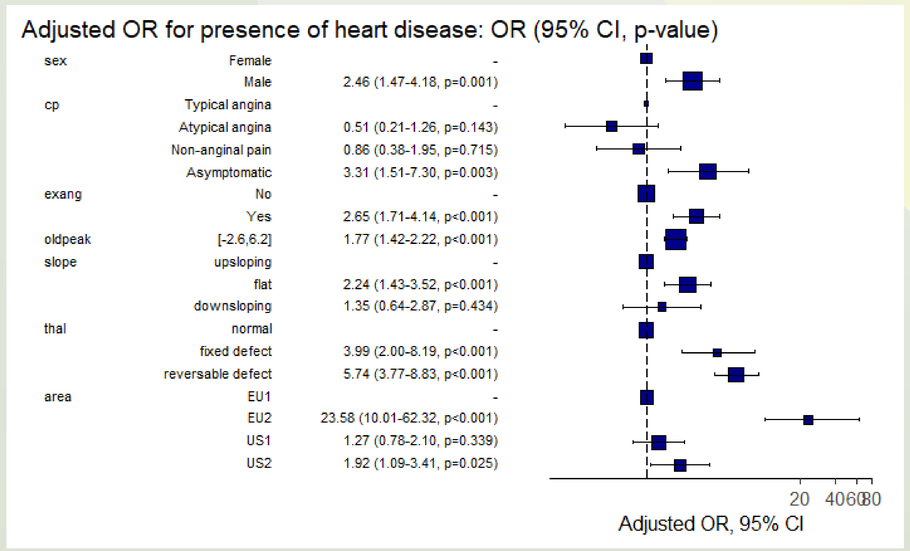
we take a look at proportions of different levels under each variable. Then the Chi-square test is utilized to check the association between the diagnosis of heart diseases and those categorial variables. We check the collinearity by exploring a generalized pairs plot.

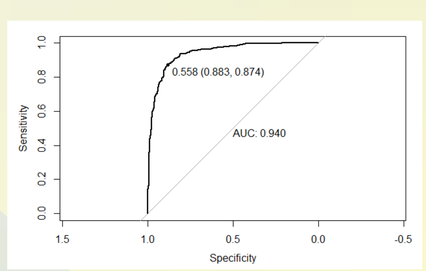
Due to the variety of variables, we choose to fit a logistic model to explore what influences the diagnosis. Beginning with the binary logistic mode, we check the unadjusted odds ratio (OR) with 95% confidence intervals and then determine some insignificant factors of interest. We continue the multivariable logistic model by excluding those irrelevant variables and checking the odds ratio with 95% confidence intervals again. ROC curve (receiver operating characteristic curve) and AUC (area under the ROC curve) are useful in evaluating the performance of the fitting model.

Cumulative link models for Ordinal Regression is introduced for estivating the relationship among ordinal severity levels of heart disease and factors of interest. We check the odds ratio with 95% confidence intervals and check if there is any insignificant variables in the model. For those variables related with the severity of the heart diseases, we then check if they satisfy the proportional hazards (PH) assumption. It is necessary to delete those variables that violate the PH assumption. By updating the model and using ANOVA to check if there is an improvement from the original model to the latest one. ROC and AUC are also applied to evaluate the performance of the model. According to these models and analysis, the combination of relative low-cost medical tasks can be explored for doctors and doctors to consider in the future.

RESULTS

After calculating some basic statistics on and taking the hypothesis tests on the continuous and categorical variables, all of variables are associated with the diagnosis of the heart disease as their p-values are all less than 0.05 significance level. The generalized pairs plot shows no collinearity among variables. We can continue the statistical analysis and modeling based on these independent variables. For the binary logistic model, we check all odds ratio with 95% confidence interval. We find that variables age, trestbps, fbs, restecg, thalach, and ca are not significant as their p-values are greater than 0.05 significance level. By excluding these variables, we modify the logistic regression model and focus on the variables, including those variables shown in the below adjusted odds ratio table and plot. By checking the ROC and AUC (the fist one below), The higher value of the upper-right corner value of ROC and the higher the AUC, the better the model is at distinguishing between patients with the disease and no disease. AUC equal to 0.94 in the above plot indicates the model works well.





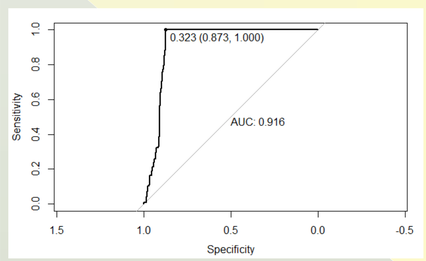


Table. Adjusted OR and CI of Cumulative Model

	Adjusted OR	95% CI
sexMale	3.23	(2.22, 4.75)
restecgST-T Abnormality	1.54	(1.11, 2.14)
restecgLeft ventricular hypertrophy	1.35	(0.96, 1.9)
thalach	0.98	(0.97, 0.98)
exangYes	2.39	(1.78, 3.2)
oldpeak	1.90	(1.67, 2.17)
slopeflat	1.85	(1.36, 2.54)
slopedownsloping	1.53	(0.96, 2.43)

Table. Intercept between cumulative levels of heart disease severity

Between cumulative category	Proportional OR
Diag 0|Diag 1	0.44
Diag 1|Diag 2	1.7
Diag 2|Diag 3	4.89
Diag 3|Diag 4	32.46

Due to the different and ordinal levels of heart disease severity, we conduct a cumulative link model to investigate the relations between 13 variables and the outcome of the disease. By first taking all the variables into consideration, the model shows the heart disease is irrelevant with age (0.284), trestbps (0.210), and fbs (0.114) as their p-values are larger than 0.05. We update the model by excluding these variables. Even though the updates model shows the validation, we need to check if every factor of interest meet the PH assumptions. The test shows that sex, restecg, thalach, exang, oldpeak and slope satisfy the PH assumptions as their p-values are greater than 0.05 significance level. However, variables including cp, thal, ca and area do not satisfy the PH assumptions as their p-values are less than 0.05 significance level. Then, we only take variables sex, restecg, thalach, exang, oldpeak and slope in to consideration for updating the model. We want to guarantee the difference between the first updated and the second updated model by checking ANOVA. The p-value less than 0.05 significance level implies the difference between two model. Thus, we go along the latest model. Finally, we achieve the fitting model and measure both the likelihood of having the heart disease and the likelihood of have different severity level of the heart disease in the above right two adjusted odds ratio table. We also check the AUC and ROC in the second plot above. The high AUC in the plot shows the good performance of the model.

CONCLUSIONS

The medical tests play an important role in helping doctors diagnosing the heart disease. Statistical analysis suggests the insignificance of medical tests in deterring the diagnosis of the heart disease including cp, thestbps, fbs, ca, thal, and thalach. Form the GLM model, we find that age, trestbps fbs restecg thalach and ca are not significant. We then exclude these variables and fit another model. The AUC is 0 94 showing the model performs well. In the CLM model, we find that these variables are significant, including sex, restecg thalach exang oldpeak and slope. Meanwhile, medical tests including restecg, exang, oldpeak, and slope play an important role in helping diagnosing the heart disease. As a result, the choice on medical tests should be restecg, exang, oldpeak, and slope tests. The total cost of these medical test is $102.8 ($15.50 for restecg test and $87.30 for exang, oldpeak, and slope tests). This cost is relatively low and acceptable even if the one does not have the insurance. We have to admit that the limitation of this study is that some variables do not meet the PH assumptions. This may affect the results and we need the further study.