Covariates for Developing Heart Diseases

WANYI HE (wah4001)

Weill Cornell Medicine

ABSTRACT

Heart disease had been recognized by World Health Organization as one of the most dangerous diseases that lead to death. An estimated number of 12 million deaths occur worldwide every year due to the heart disease and 1.4 million of the cases occurred in United States. The introducing study is an observational on heart disease that potential cardiovascular patients are sent for multiple tests, where the results will be helpful for the doctors to make diagnosis. The data were collected from two United States hospital and Two European hospital with demographic information and 11 test results.

OBJECTIVE

The primary objective of this study is to identify potential reasons that are significant to whether to develop a heart disease as well as its severity. The secondary objective is to find out the balance between the study costs and effectiveness.

METHOD

Part I. LOGISTIC REGRESSION

Consider the presence of narrowing vessel as response variable, it is reasonable to develop a logistic regression model for binary outcome. The model with p predictors is shown as below:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

The model is assuming the linearity of log odds.

Part II. MULTINOMIAL LOGISTIC REGRESSION

Consider the severity of heart disease (number of narrowing vessels), it is reasonable to bring out a multinomial logistic regression model for categorical response variable with multiple levels. Moreover, since the severity of heart disease is ordinal, the proportional odds model will be considered. The model is assuming the same effects for different levels of response variable. The model with p predictors is shown as below:

$$\log\left(\frac{P(Y \le k)}{P(Y > k)}\right) = \log\left(\frac{\gamma_k}{1-\gamma_k}\right) = \beta_{k0} + \beta_1 x_1 + \cdots + \beta_p x_p$$

$k = 1, 2, \ldots, K-1$

Part III. ASSOCIATION

To identify the associations between variables, t-test, Wilcoxon test, ANOVA test, Kruskal-Wills test, chi-square test and fisher's exact test will be considered.

RESULT

PART I. BINARY LOGISTIC REGRESSION

HEART DISEASE DIAGNOSIS	Heart disease	No heart disease	p.overall
	N=509	N=411	
AGE	55.9 (8.72)	50.5 (9.43)	< 0.001
SEX			< 0.001
Female	50 (9.82%)	144 (35.0%)	
Male	459 (90.2%)	267 (65.0%)	
CP			< 0.001
Typical angina	20 (3.93%)	26 (6.33%)	
Asymptomatic	392 (77.0%)	104 (25.3%)	
Atypical angina	24 (4.72%)	150 (36.5%)	
Non-anginal pain	73 (14.3%)	131 (31.9%)	
TRESTBPS	134 (20.5)	130 (16.8)	0.001
CHOL	177 (127)	229 (74.9)	< 0.001
FBS			< 0.001
FALSE	405 (79.6%)	367 (89.3%)	
TRUE	104 (20.4%)	44 (10.7%)	
RESTECG			0.003
Normal	284 (55.8%)	268 (65.2%)	
Having ST-T wave abnormality	119 (23.4%)	61 (14.8%)	
Left ventricular hypertrophy	106 (20.8%)	82 (20.0%)	
THALACH	127 (24.1)	148 (24.3)	< 0.001
EXANG			< 0.001
No	202 (39.7%)	350 (85.2%)	
Yes	307 (60.3%)	61 (14.8%)	
OLDPEAK	1.26 (1.19)	0.42 (0.72)	< 0.001
SLOPE			< 0.001
upsloping	109 (21.4%)	224 (54.5%)	
downsloping	75 (14.7%)	28 (6.81%)	
flat	325 (63.9%)	159 (38.7%)	
CA	1.15 (1.04)	0.30 (0.67)	< 0.001
THAL			< 0.001
normal	121 (23.8%)	314 (76.4%)	
reversable defect	388 (76.2%)	97 (23.6%)	

Table.1 logistics model with binary outcome

The descriptive statistics for binary outcome shows that **all variables are significant at a 0.05 level**. The binary logistic regression model for explaining whether to develop heart disease is displayed in the table below.

	Coefficients	OR	p-value
(Intercept)	3.9669	52.8205	< 0.0001
AGE	0.0271	1.0275	< 0.0001
SEX			
Male	0.9435	2.569	0.0005
CP			
Asymptomatic	1.0985	2.9997	0.0048
Atypical angina	-0.8	0.4493	0.0743
Non-anginal pain	-0.3174	0.728	0.438
CHOL	-0.0042	0.9958	< 0.0001
EXANG			
Yes	0.9897	2.6904	< 0.0001
OLDPEAK	0.3953	1.4848	< 0.0001
SLOPE			
Downsloping	0.2956	1.3439	0.4262
Flat	0.7728	2.1658	< 0.0001
CA	0.9434	2.5687	< 0.0001
THAL			
Reversable defect	1.5611	4.7641	< 0.0001

Table.2 logistics model with stepwise selection

Part II. ORDINAL LOGISTIC REGRESSION

	β	e^{β}	p-value
CP			
Asymptomatic	0.5886	1.8015	0.072
Atypical angina	-1.1065	0.3307	0.0047
Non-anginal pain	-0.2947	0.7448	0.3974
CHOL	-0.0014	0.9986	0.02
THALACH	-0.0138	0.9863	< 0.0001
OLDPEAK	0.5646	1.7587	< 0.0001
SLOPE			
Downsloping	0.5709	1.7699	0.019
Flat	0.5758	1.7786	0.0005
CA	0.7178	2.0499	< 0.0001
THAL			
Reversable defect	1.2311	3.425	< 0.0001

Table.3 Ordinal multinomial regression model

$$\log\left(\frac{P(diag = 0)}{P(diag > 0)}\right) = -0.4320 + \beta X$$

$$\log\left(\frac{P(diag \le 1)}{P(diag > 1)}\right) = -1.2415 + \beta X$$

$$\log\left(\frac{P(diag \le 2)}{P(diag > 2)}\right) = -2.4185 + \beta X$$

$$\log\left(\frac{P(diag \le 3)}{P(diag > 3)}\right) = -4.4257 + \beta X$$

As the severity of heart disease is regarded as ordinal categories, an ordinal logistic model is performed. Note that the model is built on proportional odds assuming parallel effects in heart disease. The four models is shown in Table.3 where intercepts based on different comparisons are listed above.

BIBLIOGRAPHY

- *Cardiovascular Disease Research*. Kathleen Hines - https://www.hopkinsmedicine.org/gim/research/content/cvd.html

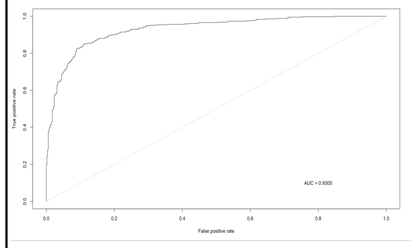
CONCLUSION

I. RELATIONSHIPS AND POTENTIAL CONFOUNDERS

	SEX	CP	TRESTBPS	CHOL	FBS	RESTECG	THALACH	EXANG	OLDPEAK	SLOPE	CA	THAL
AGE	**0.062**	< 0.001	< 0.001	**0.003**	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
SEX		< 0.001	**0.851**	< 0.001	**0.01**	**0.078**	< 0.001	< 0.001	**0.001**	< 0.001	< 0.001	< 0.001
CP			**0.14**	**0.003**	**0.095**	**0.003**	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
TRESTBPS				**0.006**	< 0.001	**0.035**	**0.003**	< 0.001	< 0.001	**0.012**	< 0.001	**0.014**
CHOL					**0.365**	< 0.001	< 0.001	**0.609**	**0.185**	**0.091**	**0.063**	**0.002**
FBS						< 0.001	**0.01**	**0.181**	**0.005**	**0.009**	< 0.001	< 0.001
RESTECG							< 0.001	**0.038**	**0.011**	**0.002**	**0.082**	**0.01**
THALACH								< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
EXANG									< 0.001	< 0.001	< 0.001	< 0.001
OLDPEAK										< 0.001	< 0.001	< 0.001
SLOPE											< 0.001	< 0.001
CA												< 0.001

Identified that most of the variables has significant association with each other, their might be potential confounders in both logistic regression and multinomial model. Multicollinearity effect might be applied to the model, which however, is hard to be identified.

II. MODEL DIAGNOSIS



Graph.1 ROC curve

By performing the ROC curve on logistics regression model, the given AUC = 0.9305 indicates a very good performance of the model on the prediction. However, it might not be accurate because of multicollinearity and variable dependencies.

III. COST EFFECTIVENESS

With the purpose of increasing the efficiency of the study, the test which might cost a lot and with respectively less significance on the outcome will be excluded. Therefore, the "slope" test will be eliminated, which costs $87.30 and needs one day laboratory work. The hypothesis on it shows a comparatively insignificance.

IV. MODEL INTERPRETATION

1. **Logistic Regression Model**

(a) $\boldsymbol{\beta_{age}}$: One-year increase in age will increase the log odds of developing heart disease by 2.75%.

(b) $\boldsymbol{\beta_{exang}}$: The log odds ratio of developing heart disease is 2.6904 for those who has exercise induced angina compare to those without.

2. **Ordinal Regression Model**

(a) $\boldsymbol{\beta_{chol}}$: The log odds ratio of will not develop heart disease compares to developing heart diseases from level 1 to level 4 severity for one mg/dl increase in serum cholestoral is 0.9986.

(b) $\boldsymbol{\beta_{thal}}$: The log odds ratio of developing heart disease that severity not greater than 2 compares to developing heart disease that severity greater than 2 is 3.425 for reversible defect compare to normal defect.