Association of clinical characteristics with heart failure

Yongsheng Li

2021-12-10

Contents

Introduction

Cardiovascular diseases (CVDs) is the NO.1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs. Most cardiovascular diseases can be prevented by addressing behavioral risk factors. In this study, we use heart-failure clinical data to explore its risk factors.

Methods

Data source

The heart-failure data source link is at data source link. This dataset consists of 299 patients with heart failure collected in 2015 with no missing data, including 8 clinical characteristics and 3 non-clinical characteristics including age(continuous), $smoking$(1=smoking,0=not),

sex(1=woman,0=not). $heart\_failure$(1=Yes, 0=No) is our dependent variable. 8 clinical characteristics are as follows:

- $anaemia$: Decrease of red blood cells or hemoglobin (boolean)
- $creatinine\_phosphokinase$: Level of the CPK enzyme in the blood (mcg/L)
- $diabetes$: If the patient has diabetes (boolean)
- $ejection\_fraction$: Percentage of blood leaving the heart at each contraction (percentage)
- $high\_blood\_pressure$: If the patient has hypertension (boolean)
- $platelets$: Platelets in the blood (kiloplatelets/mL)
- $serum\_creatinine$: Level of serum creatinine in the blood (mg/dL)
- $serum\_sodium$: Level of serum sodium in the blood (mEq/L)

Research objectives

The aim of this study is to explore the potential association of these 8 clinical characteristics with heart failure. We also want to check if age, $smoking$, sex are confounders and also if sex is a meaningful effect modifier.

Statistical analytic plan

We firstly get the descriptive statistics to explore the data. Then we do linearity check for all the variables to determine the functional form of the variable with the dependent variable. Grouped smooth, fractional polynomials and LOESS plot methods are used to assess linearity and make possible transformations when necessary.

For categorical variables in the data, we keep it unchanged. And for continuous variables, we keep it unchanged when they are linear, but convert them to categorical variables according to their quantiles or take the suggested form of fractional polynomials.

For modelling process, first, we build a univariate logistic regression on each independent clinical variable to get its basic relationship with heart failure. Second, we build an multivariate model using these 8 clinical characteristics to get the unadjusted model. Third, we check if age,$smoking$ or sex is a potential confounder. The criteria for determining a confounder is when it causes more than 10% change to original coefficients and also sensibly able to be a cause to both our independent variables of interest and dependent variable. The identified confounders will be included in the model to form the adjusted model. Fourth, we check if sex is an effect modifier(interaction term). A significant interaction term is identified when the wald test for the coefficient is statistically significant. Then, we get the final model, which includes both meaningful confounders and significant effect modifiers.

After getting the final model, we do Hosmer-Lemeshow GOF Test to evaluate the goodness of fit and do some model diagnostics to identify possible influential points. Last but not least, we report our model and give some conclusions.

Statistical criterias

All the statistical significances are determined by p-value<0.05. A rule of thumb for pseudo-R^2 between of 0.2 to 0.4 indicates excellent fit.

Code link

The complete code of this study can be found at compelte code link.

Results

Descriptive statistics

The descriptive statistics is in Table 1. The data has 299 observations with no missing data and 203 of them have heart failure. age, $creatinine\_phosphokinase$, $ejection\_fraction$, $serum\_creatinine$, $serum\_sodium$ are continuous variables and others are binary. The p-value is provided to compare the heart-failure group and non-heart-failure group for each variable, we can see some of them are significant.

Table 1: Descriptive statistics of the variables

Characteristic	N	0, N = 203[1]	1, N = 96[1]	p-value[2]
age	299	60 (50, 65)	65 (55, 75)	**<0.001**
anaemia	299	83 (41%)	46 (48%)	0.3
creatinine_phosphokinase	299	245 (109, 582)	259 (129, 582)	0.7
diabetes	299	85 (42%)	40 (42%)	>0.9
ejection_fraction	299	38 (35, 45)	30 (25, 38)	**<0.001**
high_blood_pressure	299	66 (33%)	39 (41%)	0.2
platelets	299	263,000 (219,500, 302,000)	258,500 (197,500, 311,000)	0.4
serum_creatinine	299	1.00 (0.90, 1.20)	1.30 (1.08, 1.90)	**<0.001**
serum_sodium	299	137.0 (135.5, 140.0)	135.5 (133.0, 138.2)	**<0.001**
sex	299	132 (65%)	62 (65%)	>0.9
smoking	299	66 (33%)	30 (31%)	0.8

[1]Median (IQR); n (%)

[2]Wilcoxon rank sum test; Pearson's Chi-squared test

Linearity check

The assumption for logistic regression is variable linearity. We do not need to check linearity for binary variables $anaemia$, $diabetes$ and $high\_blood\_pressure$. For other variables, we use group smooth, fractional polynomial, LOESS plot methods to assess linearity and make possible

transformations when necessary. We use fractional polynomials to determine the functional form of covariate age and do not check for $smoking$ and sex. The checking results and corresponding variable transformation results is in Table 2.

Table 2: Linearity check and variable transformation

variables	linearity	tranformation	new_variable
age	linear	-	-
anaemia	binary	-	-
creatinine_phosphokinase	linear	-	-
diabetes	binary	-	-
ejection_fraction	nonlinear	dummification from quantile of 4	ejection_fraction.q
high_blood_pressure	binary	-	-
platelets	linear	divide by 1000	platelet_kilo
serum_creatinine	nonlinear	dummification from quantile of 4	serum_creatinine.q
serum_sodium	linear	-	-
sex	binary	-	-
smoking	binary	-	-

Univariate logistic regression

We do univariate logistic regression on each of these variables to get the simple relationship of Y on each variable X. The results are in Table 3. We can see age, $ejection\_fraction.q$, $serum\_creatinine.q$, $serum\_sodium$ are significant, to which we need to pay attention.

Table 3: simple univariate relationship of Y with each X

Characteristic	N	Event N	OR[1]	95% CI[1]	p-value
age	299	96	1.05	1.03, 1.07	**<0.001**
anaemia	299	96			0.25
0			—	—	
1			1.33	0.82, 2.17	
creatinine_phosphokinase	299	96	1.00	1.00, 1.00	0.29

[1]OR = Odds Ratio, CI = Confidence Interval

Characteristic	N	Event N	OR[1]	95% CI[1]	p-value
diabetes	299	96			0.97
0			—	—	
1			0.99	0.60, 1.62	
ejection_fraction.q	299	96			**<0.001**
[14,30]			—	—	
(30,38]			0.27	0.14, 0.50	
(38,45]			0.15	0.06, 0.34	
(45,80]			0.25	0.12, 0.51	
high_blood_pressure	299	96			0.17
0			—	—	
1			1.42	0.86, 2.34	
platelets_kilo	299	96	1.00	1.00, 1.00	0.39
serum_creatinine.q	299	96			**<0.001**
[0.5,0.9]			—	—	
(0.9,1.1]			3.31	1.47, 8.04	
(1.1,1.4]			3.13	1.33, 7.86	
(1.4,9.4]			13.3	5.98, 32.6	
serum_sodium	299	96	0.91	0.85, 0.96	**<0.001**
sex	299	96			0.94
0			—	—	
1			0.98	0.59, 1.64	
smoking	299	96	0.94	0.56, 1.58	0.83

[1]OR = Odds Ratio, CI = Confidence Interval

Confounding checking

We found that age and sex can cause many coefficients to change more than 10%, whereas $smoking$ causes almost no changes to the coefficients. In addition, both age and sex are sensibly causes for the clinical characterstics and for risk of heart failure. Therefore, we will include age and sex as meaningful confounders into model. Note that age can even cause the coefficient of diabetes to have a sign change.

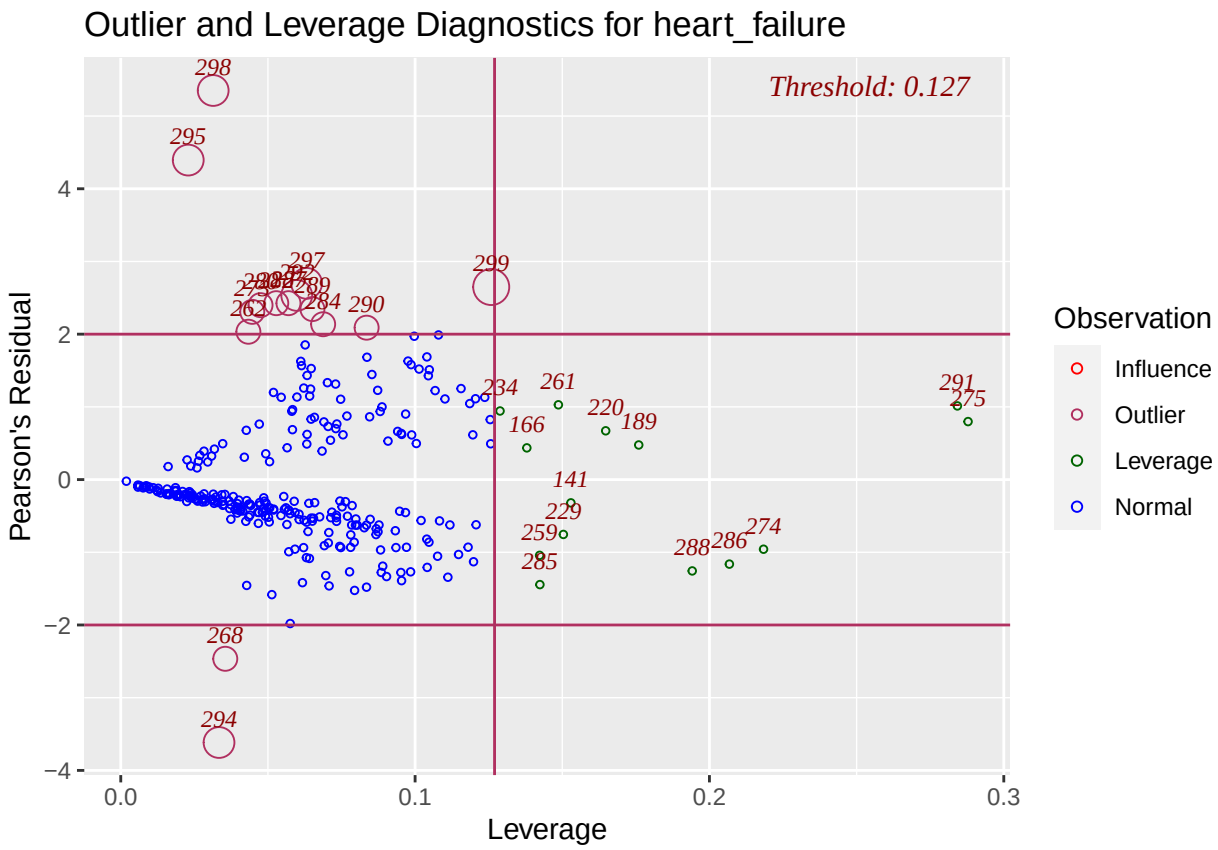
Table 4: Parameter changes after incorporating potential confounders

coefficients_change	age(%)	smoking(%)	sex(%)
anaemia1	-4.19	0.24	-4.84
creatinine_phosphokinase	**19.03**	0.05	1.97
diabetes1	**-258.35**	-1.75	**47.73**
ejection_fraction.q(30,38]	**23.04**	0.01	0.32
ejection_fraction.q(38,45]	**16.52**	-0.05	1.30
ejection_fraction.q(45,80]	**30.48**	-0.12	4.10
high_blood_pressure1	**-10.04**	0.09	-3.74
platelets_kilo	4.46	0.02	**14.28**
serum_creatinine.q(0.9,1.1]	**-18.39**	-0.04	3.04
serum_creatinine.q(1.1,1.4]	**-36.96**	-0.07	0.13
serum_creatinine.q(1.4,9.4]	**-14.49**	0.02	0.57
serum_sodium	7.46	-0.03	2.76

Interaction checking

With age and sex included as confounders, we get the adjusted model. We further check if sex is a significant effect modifier. After putting interaction term of sex with each of these 8 clinical characteristics into the adjusted model respectively, the wald test result of each interaction term shows that sex interacts with $platelets\_kilo$(z-statistic=-2.143, p-value=0.032140) and $serum\_creatinine.q$(z-statistic=2.046, pvalue=0.040746 for (0.9,1.1]; z-statistic=1.490, p-value=0.136233 for (1.1,1.4]; z-statistic=4.497, pvalue=6.91e-06 for (1.4,9.4]) significantly on the risk of heart failure. Therefore we will include these interaction terms to get the final model.

Model goodness of fit and diagnostics



The pseudo-R^2 is 0.2887255, indicating the final model is rather well. The Hosmer-Lemeshow GOF Test shows no significant departure from goodness of fit(statistic=17.576491, p-value=0.4838621). In addition, we can see there are no influential points based the Pearson's-Leverage plot from the final model .

Conclusion

The final model is as follows:

Table 5: Results of the final multivariable model

Characteristic	OR[1]	95% CI[1]	p-value
anaemia			
0	—	—	
1	1.48	0.77, 2.83	0.2
creatinine_phosphokinase	1.00	1.00, 1.00	**0.032**
diabetes			

[1]OR = Odds Ratio, CI = Confidence Interval

Characteristic	OR[1]	95% CI[1]	p-value
0	—	—	
1	1.00	0.53, 1.91	>0.9
ejection_fraction.q			
[14,30]	—	—	
(30,38]	0.23	0.10, 0.48	**<0.001**
(38,45]	0.11	0.04, 0.27	**<0.001**
(45,80]	0.17	0.07, 0.42	**<0.001**
high_blood_pressure			
0	—	—	
1	1.65	0.87, 3.14	0.12
platelets_kilo	0.99	0.99, 1.00	**0.022**
sex			
0	—	—	
1	0.31	0.03, 4.43	0.4
serum_creatinine.q			
[0.5,0.9]	—	—	
(0.9,1.1]	7.40	1.39, 59.9	**0.030**
(1.1,1.4]	4.73	0.81, 40.4	0.11
(1.4,9.4]	47.4	8.46, 428	**<0.001**
serum_sodium	0.96	0.89, 1.03	0.3
age	1.06	1.03, 1.09	**<0.001**
platelets_kilo * sex			
platelets_kilo * 1	1.01	1.00, 1.02	**0.029**
sex * serum_creatinine.q			
1 * (0.9,1.1]	0.21	0.02, 1.60	0.2
1 * (1.1,1.4]	0.31	0.03, 2.59	0.3
1 * (1.4,9.4]	0.09	0.01, 0.67	**0.027**

[1]OR = Odds Ratio, CI = Confidence Interval

Our final multivariable logistic regression model included $anaemia$ vs non-anaemia,$creatinine\_phosphokinase$, $diabetes$ vs no-diabetes, $ejection\_fraction$ quantile(4 categories), $high\_blood\_pressure$ vs non-high_blood_pressure, $platelets$(in kilos), sex(woman vs man), $serum\_creatinine$ quantile(4 categories), $serum\_sodium$, age, $platelets * sex$ interactions, $serum\_creatinine.q * sex$ interactions.

We found that wald tests for $anaemia$(p-value=0.2), $diabetes$(p-value>0.9), $high\_blood\_pressure$(p-value=0.12), $serum\_sodium$(p-value=0.3) are all non-significant. Therefore we can say they have no effect on the risk of heart failure. The interesting result is that $creatinine\_phosphokinase$ has a p-value of 0.032 but OR=1, indicating $creatinine\_phosphokinase$ has almost zero effect on the risk of heart failure.

We found that those in higher ejection fraction quantile categories have lower odds of heart failure compared to those in category [14,30]: OR=0.23, 95%CI=[0.10, 0.48] for (30,38]; OR=0.11, 95%CI=[0.04, 0.27] for category (38,45]; OR=0.17, 95%CI=[0.07, 0.42] for (45,80] respectively, with all p-values<0.001. As a confounder, one unit increase in age will cause 1.06 times as likely as before to have heart failure(OR=1.06, 95%CI=[1.03, 1.09],p-value<0.001). However, sex do not have statistically significant effect on risk of heart failure.

For woman($sex = 1$), per 1000 unit increase in platelets will make it 1.001169 times as likely as before to have heart failure(calculated from raw coefficients). For man($sex = 0$), per 1000 unit increase in platelets will make it 99% as likely as before to have heart failure. Note that this is only per 1000 unit increase, and platelets can vary in a broader range to cause a larger effect on risk of heart failure.

When $sex = 0$, namely for man, those in higher serum creatinine quantile categories have higher odds of heart failure than those in category[0.5,0.9]: OR=7.40, 95%CI=[1.39, 59.9], p-value=0.030 for (0.9,1.1]; OR=4.73, 95%CI=[0.81, 40.4],p-value=0.11 for (1.1,1.4], but non-significant; OR=47.4, 95%CI=[8.46, 428], pvalue<0.001 for (1.4,9.4]. However, for woman these odds ratios are completely different: OR=1.56,1.45, 4.06, respectively for (0.9,1.1], (1.1,1.4], (1.4,9.4] compared to [0.5,0.9], calculated from original coefficients. This is a very interesting result. It means that serum creatinine will generally introduce more risk of heart failure for man than for woman(both compared to the [0.5,0.9] base category). Figure 1 shows the details of the interaction.

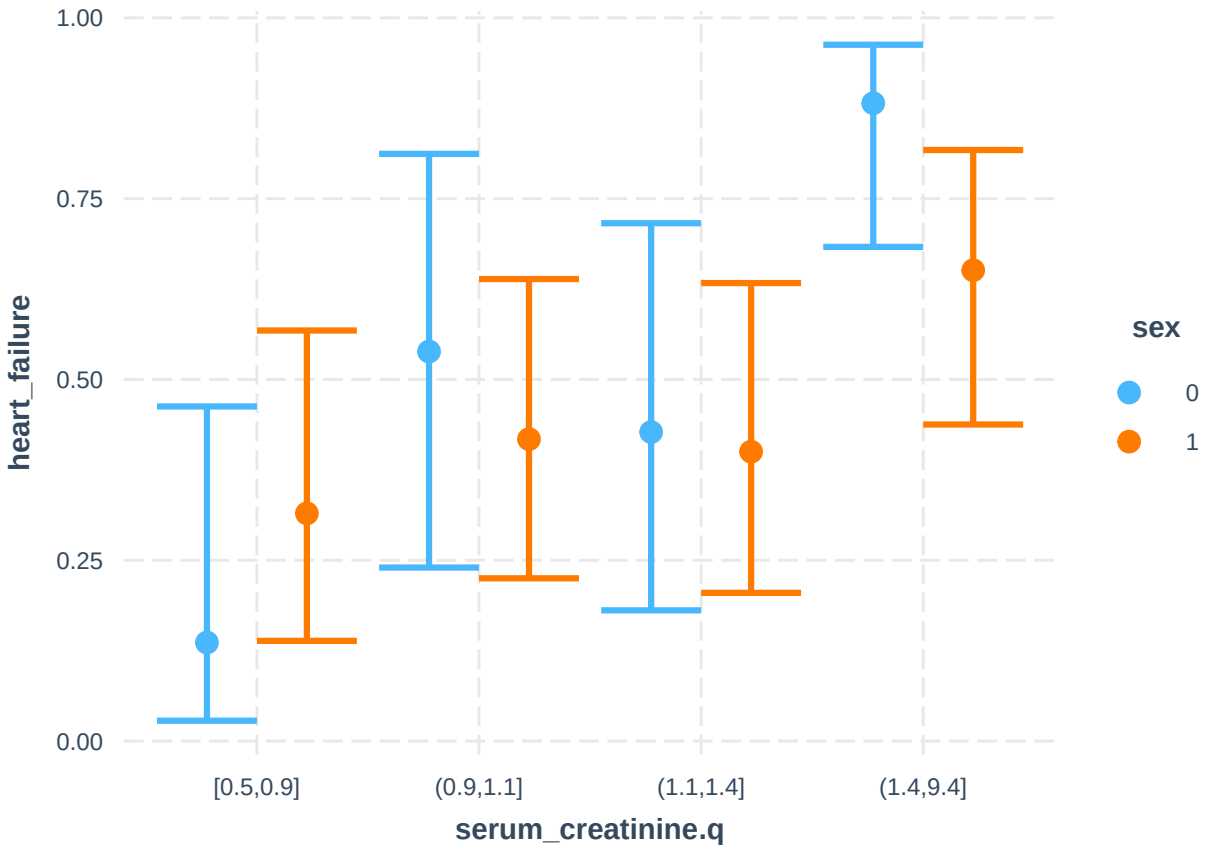


Figure 1: sex and serum_creatinine interaction on the risk of heart failure