Survival Analysis of Heart Failure

Yonghan Qiu
Mengzhao Xu
Yuhua Liu

1. Introduction

Heart failure (HF) is a clinical syndrome, occurs when your heart muscle doesn't pump blood as well as it should.[1] And it is becoming a major public health problem in recent decades, due to its increasing mortality rate. Heart failure affects nearly a million people across the UK. It's a life-limiting condition that too often causes emergency hospital admissions, poor quality of life and ultimately early death. [2] Survival analysis was applied to differ the importance of risk factors that contribute to the mortality of heart failure patients. The data of the study is based on real hospital administrative data for England called Hospital Episodes Statistics. Our simulated extract contains a random sample of emergency (unplanned) admissions for heart failure (ICD10 code I50). Many of the fields are comorbidities coded as 0/1, where 1 indicates that the patient had it recorded. The potential risk factors are los (hospital length of stay in nights), age, gender, cabg (previous heart bypass), diabetes, hypertension, ihd (ischaemic heart disease), arrhythmias, obesity, and so on, totally 18 variables. At first, we performed a preliminary analysis to get the baseline characteristics of the data and test the correlation among the risk factors additionally. Then we fit the risk factors into a Cox PH model to get a basic idea of the significance of variables. And Kaplan-Meier plots showed a gradually decreasing pattern on survival rates based on different factors during the study period. Moreover, AIC forward selection method was used to select the best Cox Regression model. At last, the Cox-Snell residual figure showed that the Cox model fit well. In conclusion, we found that los, age, gender, ihd and prior_dnas(number of outpatient appointments missed in the previous year) are the significant risk factors for mortality among heart failure patients.

2.Methodology

2.1 Data Description

The study was based on 1000 heart failure patients with 452 female and 548 male. Their ages are between 29 and 101, and 492 of them died during the study. The original dataset contains 31 various parameters, but we only take 18 of them in our study since some of them either have a little relationship with heart failure obviously or have little samples. The remaining 18 parameters are: death (0/1), los (hospital length of stay in nights), age (in years), gender (1=male, 2=female), cabg (previous heart bypass), diabetes (any type), hypertension, ihd (ischaemic heart disease), arrhythmias, copd (chronic obstructive lung disease), obesity, pvd (peripheral vascular disease), valvular_disease (disease of the heart valves), pacemaker, prior_appts_attended (number of outpatient appointments attended in the previous year), prior_dnas (number of outpatient appointments missed in the previous year), pci (percutaneous coronary intervention) and fu_time (follow-up time, i.e. time in days since admission to hospital). We changed the original variable gender(1=male, 2=female) to gender(0=male, 1=female) to make it binary.

2.2 Statistic Techniques

Survival Analysis was used to estimate the survival and mortality rates since the presence of censored data. In our dataset, 0 denotes censored data and 1 denotes death under the variable death.

The Cox-proportional hazards model was used to research the relationships between the time to event outcome and a set of explanatory variables and also test for the significance of these factors. The model has the form:

$$h(t|Z) = h_0(t)\ exp\ (\beta Z)$$

where $h_0(t)$ is the baseline hazard at time t; Z is the vector of explanatory variables; β is the vector of coefficients corresponding to the variables. To estimate and compare the survival function S(t) for different levels of explanatory variables, we used the Kaplan-Meier (K-M) estimator. The definition of K-M estimator $\hat{S}(t)$ is:

$$\hat{S}(t) = \Pi_{t_j \le t}(1 - d_j/Y_j), t_1 \le t$$

where d_j is the number of deaths that occur at t_j , Y_j is the number at risk (alive and under observation just before t_j) and $t_1 < t_2 < \cdots < t_k$ denote distinct times at which deaths occur.

We also used forward AIC to select the best Cox Regression model and plotted the Cox-Snell residual to check the overall fit of the model. The definition of the Cox-Snell residual is:

$$r_j = \widehat{H_0}(T_j)\ exp\ (\hat{\beta}^T Z_j), j = 1, \cdots, n.$$

where $Z_j = (Z_{j1}, \cdots, Z_{jp})^T$ are all fixed-time covariates. r_j are censored sample from exponential distribution, given the assumed Cox model holds and $\hat{\beta}$, $\widehat{H_0}(t)$ close to the true values β, $H_0(t)$. So $\widehat{H_r}(x)$ is Nelson-Aalen estimator of CMHF of r_j based on $\{r_j, \delta_j\}$. We also plotted $\widehat{H_r}(r_j)$ versus r_j. If the Cox model provides a good fit of the data, we expect a straight line through the origin with slope 1.

3. Analysis and Results

3.1 Preliminary Analysis

(1) Baseline Characteristics of the Data

The baseline characteristics module is designed to summarize important attributes of the participants enrolled at the start of or the baseline of the study. The baseline characteristics module itself then is composed of a table.

The following Table.1 demonstrates the mean and standard deviation of the continuous variables of dead and censored patients at the end of the follow up period, Table.2 demonstrates a count and the central tendency and dispersion for each category of dead and censored patients at the end of the follow up period.

From these tables, we can see the number of dead cases and the censored cases are similar. The continuous variables age, los and the categorical variable cabg, obesity and pci have large influence on the result obviously since the mean of age and los of dead cases have larger differences with the censored cases and the percentage of dead cases of whether have cabg, obesity, pci or not is obviously different.

Table 1. Baseline Characteristics for Dead and Censored Patients for Continuous Variables

Continuous Variables		
Variable	**Dead(N=492)**	**Censored(N=508)**
	Mean (Standard Deviation)	
age	82.175(8.788)	75.396(12.119)

los	12.447(14.475)	9.154(10.180)
prior_appts_attended	5.283(6.271)	5.785(7.344)
prior_dnas	0.547(1.223)	0.457(0.985)

note:
los: hospital length of stay in nights
prior_appts_attended:number of outpatient appointments attended in the previous year
prior_dnas; number of outpatient appointments missed in the previous year

Table 2. Baseline Characteristics for Dead and Censored Patients for Categorical Variables

Categorical Variables				
Variable	**Categories**	**Dead(N=492)**	**Censored(N=508)**	**Percentage of Dead (49.2%)**
gender	Female(1)	224(45.5%)	228(44.9%)	49.6%
	Male(0)	268(54.5%)	280(55.1%)	48.9%
cabg	Yes(1)	1(0.2%)	13(2.6%)	7.1%
	No(0)	491(99.8%)	495(97.4%)	49.8%
diabetes	Yes(1)	129(26.2%)	154(30.3%)	45.6%
	No(0)	363(73.8%)	354(69.7%)	50.6%
hypertension	Yes(1)	300(61.0%)	321(63.2%)	48.3%
	No(0)	192(39.0%)	187(36.8%)	50.7%
ihd	Yes(1)	253(51.4%)	242(47.6%)	51.1%
	No(0)	239(48.6%)	266(52.4%)	47.3%
arrhythmias	Yes(1)	231(47.0%)	259(51.0%)	47.1%
	No(0)	261(53.0%)	249(49.0%)	51.2%
copd	Yes(1)	127(25.8%)	115(22.6%)	52.5%
	No(0)	365(74.2%)	393(77.4%)	48.2%
obesity	Yes(1)	23(4.7%)	35(6.9%)	39.7%
	No(0)	469(95.3%)	473(93.1%)	49.8%
pvd	Yes(1)	46(9.4%)	54(10.6%)	46.0%
	No(0)	446(90.6%)	454(89.4%)	49.6%
valvular_disease	Yes(1)	116(23.6%)	128(25.2%)	47.5%
	No(0)	376(76.4%)	380(74.8%)	49.7%
pacemaker	Yes(1)	18(3.7%)	19(3.7%)	48.6%
	No(0)	474(96.3%)	489(96.3%)	49.2%
pci	Yes(1)	10(2.0%)	19(3.7%)	34.5%
	No(0)	482(98.0%)	489(96.3%)	49.6%

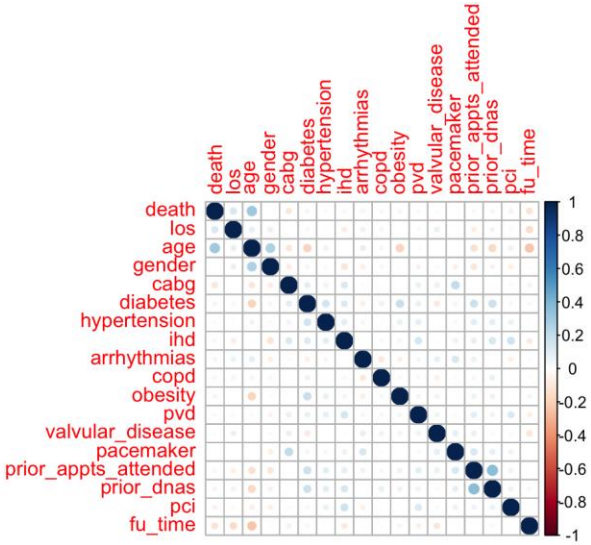
note:

cabg: previous heart bypass.
ihd: ischaemic heart disease
copd: chronic obstructive lung disease
pci: percutaneous coronary intervention

(2) Correlation

In addition, we test the correlation pairwise. From the result below we can see there are no dark colors between different variables, which means they don't have high correlation, so we can hold all variables.

Figure 1. Correlation Plot



3.2 Analysis and Results

3.2.1 Survival Function and Cumulative Hazard Function

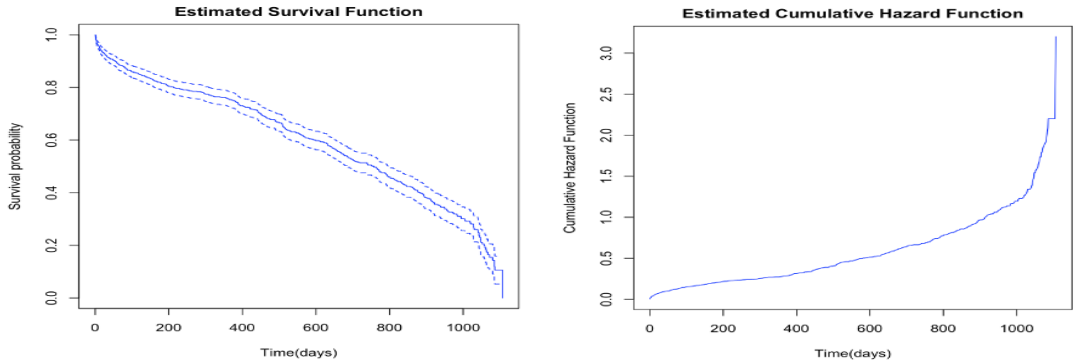
Prior to fitting a Cox PH model, we wanted to study the survival probability and cumulative hazard rate of death time in the dataset. We used K-M estimator and N-A estimator respectively. Figure 2 shows the decreasing pattern on survival probability and the increasing pattern on cumulative hazard rate. Table 3 shows their values and standard errors.

We can also see that $\hat{s}(748) = 0.5004 > 0.5$ and $\hat{s}(749) = 0.4984 < 0.5$, so the median time $\hat{x}_{0.5} = 749$ days.

Table 3. Survival Probability and Cumulative Hazard Rate

time	n.risk	n.event	surv	std.surv	cumhaz	std.chaz
0	1000	3	0.9970	0.00173	0.0030	0.00173
1	992	9	0.9880	0.00350	0.0121	0.00349
2	973	7	0.9808	0.00444	0.0193	0.00442
3	963	5	0.9758	0.00501	0.0245	0.00499
……						
748	246	0	0.5004	0.03790	0.6911	0.03783
749	245	1	0.4984	0.03812	0.6951	0.03805
……						
1104	2	0	0.0815	0.37498	2.4324	0.33947
1107	1	1	0.0000	Inf	3.4324	1.05605

Figure 2. Survival Probability Curve and Cumulative Hazard Rate Curve



3.2.2 Cox Model

We used all the variables to fit the Cox model. Table.4 presented the results of the Cox model. Age, los, gender, ihd and prior_dnas were found to be significant variables with p-value smaller than 0.05, and other variables were non-significant variables.

According to Table.4, age was the most significant variable. The results indicated that death rates increased by approximately 6.4% for each year of age. The second and third significant variables were los(hospital length of stay in nights) and prior_dnas(number of outpatient appointments missed in the previous year) , the result indicated that the death rates increased by approximately 1.4% for each night stay in hospital and 14.4% for each missed outpatient appointment in the previous year. Besides, the patients' gender and whether they had ihd(ischaemic heart disease) or not influenced the final result.

Table. 4 Significance of variables under Cox regression

Variable	β-coef	exp(coef)	se(coef)	Z-value	p-value
los	0.013	1.014	0.003	4.120	3.7e-05
age	0.062	1.064	0.006	10.760	< 2e-16
gender	-0.283	0.754	0.096	-2.940	0.003
cabg	-1.839	0.159	1.007	-1.830	0.068
diabetes	-0.012	0.988	0.113	-0.100	0.918
hypertension	-0.034	0.967	0.096	-0.360	0.722
ihd	0.245	1.278	0.096	2.540	0.011
arrhythmias	-0.152	0.859	0.095	-1.610	0.108
copd	0.095	1.100	0.105	0.900	0.366
obesity	0.108	1.114	0.224	0.480	0.629
pvd	0.046	1.047	0.161	0.290	0.774
valvular_disease	0.191	1.211	0.109	1.760	0.079
pacemaker	0.130	1.139	0.254	0.510	0.609
prior_appts_attended	-0.007	0.993	0.008	-0.850	0.393
prior_dnas	0.134	1.144	0.039	3.420	0.001
pci	-0.182	0.834	0.327	-0.560	0.579

3.2.3 Model Selection

(1) Forward Stepwise Selection

We used a forward stepwise selection method for the Cox-proportional hazards model to find the best model for the heart failure data. As shown in Table.5, AIC value was used to carry out the model fitting process. The method begins with a model that contains no variables, then starts adding the lowest AIC variables one after the other until all the variables under consideration are included in the model.

The process of forward stepwise selection method is shown in Table. 5. We have 8 risk factors for the final model: age, los, prior_dnas, gender, ihd, cabg, arrhythmias and valvular_disease. Table. 6 shows the parameter estimates for the final model. From the table, we get our final model as

$$h(t|age, los, prior\_dnas, gender, ihd, cabg, arrhythmias\ and\ alvular\_disease) = h_0(t)\exp(0.06195 \cdot age + 0.01355 \cdot los + 0.12041 \cdot prior\_dnas - 0.28157 \cdot gender + 0.23982 \cdot ihd - 1.81423 \cdot cabg - 0.16183 \cdot arrhythmias + 0.18865 \cdot valvular\_disease)$$

Compared to the Cox model without variable selection, the final model has 8 variables, including 5 variables with p value lower than 0.05 and 3 variables that are not. The first 5 variables are the same as the Cox model without variable selection. Table 6 shows the similar results to Cox model without variable selection, and just the values are slightly different.

Table. 5 AIC Forward Selection for Cox Model

Start: AIC=5914		**Step: AIC=5778**		**Step: AIC=5765**	
Surv(fu_time, death) ~ 1		**Survfu_time, death) ~ age**		**Surv(fu_time, death) ~ age + los**	
Variables	AIC	Variables	AIC	Variables	AIC
age	5778	los	5765	prior_dnas	5755
los	5891	prior_dnas	5767	ihd	5757
cabg	5905	gender	5772	gender	5758
ihd	5912	ihd	5774	cabg	5762
prior_dnas	5913	cabg	5776	valvular_disease	5764
valvular_disease	5914	valvular_disease	5777	arrhythmias	5765
obesity	5914	none>	5778	none>	5765
none>	5914	arrhythmias	5779	copd	5765
diabetes	5914	copd	5779	pvd	5766
pci	5915	pvd	5780	prior_appts_attended	5766
prior_appts_attended	5915	diabetes	5780	hypertension	5766
copd	5915	hypertension	5780	pacemaker	5767
arrhythmias	5916	prior_appts_attended	5780	diabetes	5767
pvd	5916	obesity	5780	pci	5767
gender	5916	pacemaker	5780	obesity	5767
pacemaker	5916	pci	5780		
hypertension	5916				
Step: AIC=5755		**Step: AIC=5748**		**Step: AIC=5745**	
Surv(fu_time, death) ~ age + los + prior_dnas		**Surv(fu_time, death) ~ age + los + prior_dnas + gender**		**Surv(fu_time, death) ~ age + los + prior_dnas + gender + ihd**	
Variables	AIC	Variables	AIC	Variables	AIC
gender	5748	ihd	5745	cabg	5741

ihd	5750	cabg	5745	arrhythmias	5744
cabg	5753	arrhythmias	5747	valvular_disease	5744
valvular_disease	5755	valvular_disease	5748	none>	5745
arrhythmias	5755	none>	5748	copd	5746
none>	5755	copd	5749	prior_appts_attend	5746
copd	5756	prior_appts_attended	5750	hypertension	5747
pvd	5757	pvd	5750	obesity	5747
prior_appts_attended	5757	hypertension	5750	pci	5747
hypertension	5757	pacemaker	5750	pvd	5747
obesity	5757	obesity	5750	pacemaker	5747
diabetes	5757	pci	5750	diabetes	5747
pacemaker	5757	diabetes	5750		
pci	5757				
Step: AIC=5741		**Step: AIC=5740**		**Step: AIC=5739**	
Surv(fu_time, death) ~ age + los + prior_dnas + gender + ihd + cabg		**Surv(fu_time, death) ~ age + los + prior_dnas + gender + ihd +cabg +** arrhythmias		**Surv(fu_time, death) ~ age + los + prior_dnas + gender + ihd + cabg + arrhythmias + valvular_disease**	
Variables	AIC	Variables	AIC	Variables	AIC
arrhythmias	5740	valvular_disease	5739	none>	5739
valvular_disease	5740	none>	5740	copd	5740
none>	5741	copd	5741	prior_appts_attended	5740
copd	5742	prior_appts_attended	5741	pci	5740
prior_appts_attended	5742	pci	5741	obesity	5740
obesity	5743	obesity	5741	hypertension	5740
pci	5743	pacemaker	5741	pacemaker	5740
hypertension	5743	hypertension	5741	pvd	5741
pvd	5743	diabetes	5741	diabetes	5741
diabetes	5743	pvd	5741		
pacemaker	5743				

Table. 6 Parameter Estimates for the Final Cox Model

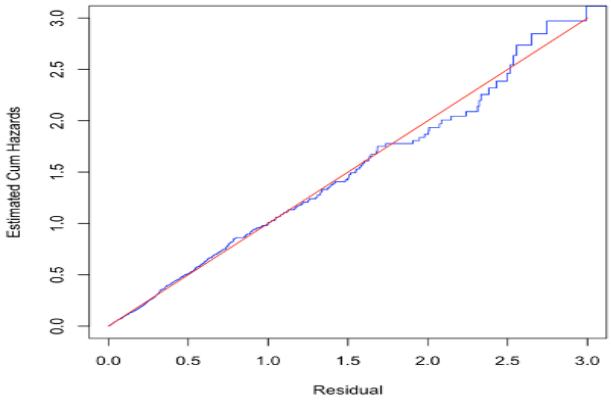
Variables	coef	exp(coef)	se(coef)	Z-value	P-value
age	0.062	1.064	0.006	11.040	< 2e-16
los	0.014	1.014	0.003	4.290	1.8e-05
prior_dnas	0.120	1.128	0.036	3.380	0.001
gender	-0.282	0.755	0.095	-2.960	0.003
ihd	0.240	1.271	0.093	2.580	0.010

cabg	-1.814	0.163	1.004	-1.810	0.071
arrhythmias	-0.162	0.851	0.091	-1.770	0.076
valvular_disease	0.189	1.208	0.107	1.760	0.078

(2) Cox-Snell Residual Plot

As in Figure 3, we conducted a Cox-Snell Residual plot to assess the fitness of our final Cox model. From the plot, we see that the estimated cumulative hazards follow closely to the 45-degree straight line. Therefore, we can conclude that the model is a good fit.

Figure 3. Cox-Snell Residual Plot



3.2.4 Highly Correlated Features

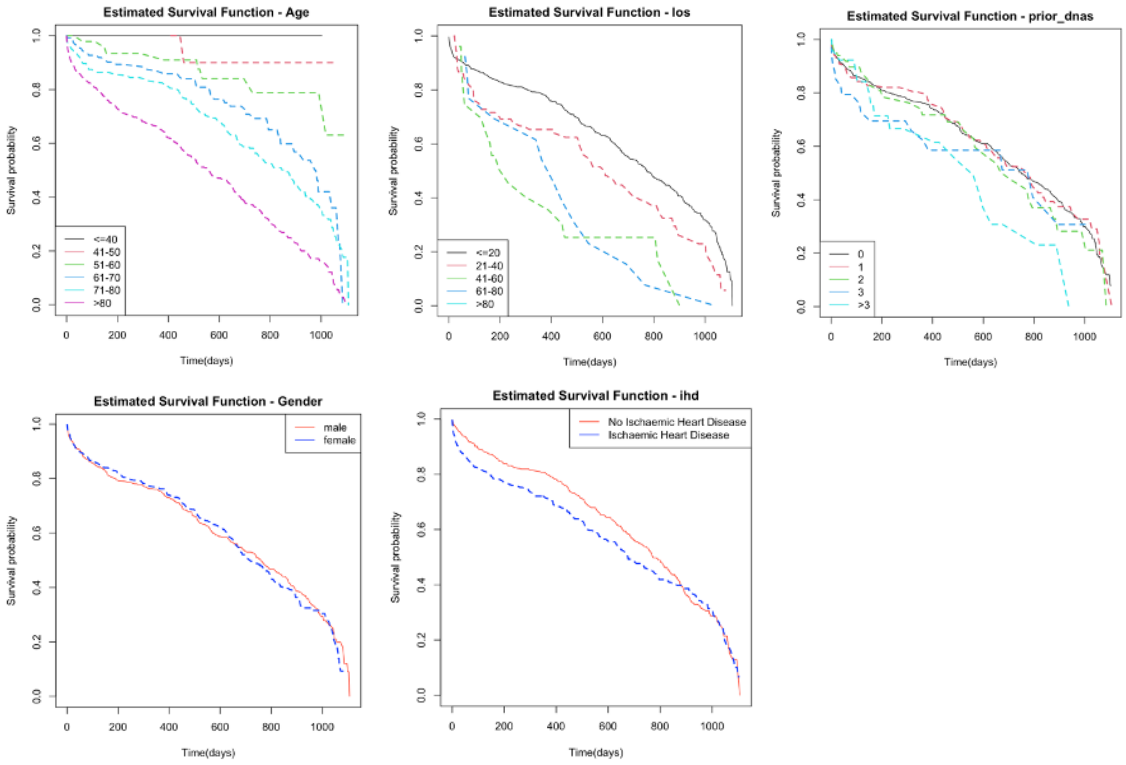
Now let's focus on the highly correlated features with p-values of <0.05 from the Cox PH models with and without variable selection. These are the ones that show high correlation with the mortality rate. There are 5 highly correlated variables, including age, los, prior_dnas, gender and ihd. We studied the survival probability under different correlated variables. For the first 3 continuous variables, we divided them into different groups.

Figure 4 shows survival functions for different covariables. From the pictures we can see

- For age, splitting patients up by age groups shows a large difference between each age group. In particular, patients younger than 40 have a survival probability of 1.
- For los, each group also has an obvious difference, but not as large as age, which is the same as Table 5, as exp(coef) of age is 1.064 and exp(coef) of los is 1.014.
- For prior_dnas, when the number of outpatient appointments missed is less than 3, the survival curves don't have a significant difference, but if the number is more than 3, the survival probability is significantly lower.

- For gender, the differences between the two curves here are not obvious, but the p-value=0.003 tells a different story. This is why eyeballing the survival curves alone is not sufficient to determine feature correlation.
- For ihd, the two curves here are obviously different, and patients who don't have ischaemic heart disease have a higher survival probability.

Figure 4. Survival Function for Different Covariables



4. Conclusion

In this study, firstly Cox regression was used to model hazard rate for the heart failure patients, and the results showed that age, los, prior_dnas, gender and ihd were identified as significant variables. Secondly, using AIC forward selection method, we obtained the best model, including 8 variables (age, los, prior_dnas, gender, ihd, cabg, arrhythmias and valvular_disease), and the first 5 variables were significant, which was the same as the Cox model without variable selection. Finally, we used Cox-Snell Residuals to test whether the results fit our model or not, and the answer is yes.

The outputs can be concluded that the patients with high values of los (greater than 61) and high values of prior_dnas (greater than 3) have higher death risks. What's more, the death rate increases with growing age. In addition, ihd (ischaemic heart disease) has a significant influence on mortality rate. Even no significant differences in the plots were found between gender for death risks, the p-value of gender showed it was significant actually.

5. Reference

Data from https://www.kaggle.com/datasets/jackleenrasmybareh/heart-failure

[1] Mayo Clinic. Heart failure. Retrieved from https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142

[2] British Heart Foundation(2020). Heart failure: A blueprint for change.