ISSN 2278-3091
Volume 9, No.1.4, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse9891.42020.pdf
https://doi.org/10.30534/ijatcse/2020/9891.42020

A Comparison of Parametric Survival Models to Predict Risk Factors towards Hemodialysis Patient

Nur Aini Ruslan[1], Nik Arni Nik Mohammad[2] & Nor Azura Md.Ghani[3]
[1,2,3]Center for Statistical and Decision Sciences Studies, Faculty of Computer & Mathematical Sciences Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[3]National Design Centre Universiti Teknologi MARA 40450 Shah Alam, Selangor, Malaysia

ABSTRACT

Kidney failure, also called end-stage-renal disease (ESRD), is the last stage of chronic kidney disease where kidney functions only works at less than 15% of the normal capacity that lead to hemodialysis. Hemodialysis (HD) is the most common renal replacement therapy for patients with ESRD. HD patients have high mortality risk, influenced by the risk factors which increased the probability of patients dying. The main purpose of this study is to identify the best survival model to determine risk factors that lead to mortality among HD patients using survival analysis. A retrospective cohort study was conducted at five HD centers in Kota Bharu, Kelantan, involving 105 HD patients, between Oct 1,2005 to Apr 30,2017. Standard parametric survival analysis was used in this study after model adequacy checking was done using the Kaplan-Meier curves. Exponential, Weibull, log-logistic and log-normal distributions were used to obtain the best parametric survival model, where the Akaike information criterion (AIC), Bayesian information criterion (BIC) and log-likelihood value were used to compare the best performance among parametric survival models. The results show that 72.29% of the patients were alive at the end of this study, while 25.71% died. The median survival time for a HD patient in 4229 days of follow-up period is 2173 days. The Weibull distribution was chosen as the best model to identify the risk factors that affect the survival of HD patients because this distribution had the lowest AIC and BIC, and the highest log-likelihood value. It was found that diabetes mellitus (p-value = 0.032) is the only significant risk factor that lead to mortality among HD patients

Key words : Hemodialysis, Kidney Disease, Parametric Survival Models, Risk Factors

1. INTRODUCTION

Chronic kidney disease explains the gradual loss of kidney function. It is the last stage of chronic kidney disease, thus called end-stage renal disease (ESRD). Patients are at higher risk of kidney failure when he/she suffers from diabetes, high blood pressure/hypertension and smoking. High prevalence of cardiovascular risk factors such as hypertension, diabetes mellitus, smoking and advanced age had a high risk of mortality in ESRD patients, as stated by do [1]. HD patients who smoke was less likely to receive a kidney transplant and more likely to die than non-smokers.

Hemodialysis (HD) is a procedure that is performed routinely on persons who suffer from ESRD, blood is filtered outside of the body. The filtered blood then flows back to the body. ESRD patients suffer from very high mortality. Reference [2] found that 6% of HD patients died within the first 90 days.

Comorbidities are the coexistence of two or more disorders or illnesses in a patient. In a study by [3], he reported that those with at least three comorbidities had three times higher risk of mortality compared to patients with no comorbidities. The most common comorbidities among patients with ESRD are hypertension, diabetes mellitus, anemia and gout. Women, in general, experience more life expectancy compared to men. Reference [4] stated that life expectancy was longer for women and patients diagnosed with kidney failure undergoing HD are mostly male.

Body mass index (BMI) is described as references for weight control in the general population. It defines the height and weight characteristics classified according to groups. Reference [5] concluded that the worst survival occurs in patients of the lowest two quartiles of BMI. Low BMI as an independent risk factor for mortality in HD patients had been studied. In a study of 9714 HD patients by [6], BMI<30 have higher mortality rate compared to patients with BMI≥30 kg/m^2.

Gout, as stated by [7], is a major prevalent form of inflammatory arthritis. Gout is really connected to chronic kidney disease. People who suffered from chronic kidney disease might get the first attack of gout when kidney function progressively declines due to reduced urinary excretion of uric acid.

2. METHOD

Survival analysis is commonly known as the analysis of time-to-event data. The data describes the length of time from a time of origin to an endpoint of interest [8]. According to [9], survival analysis is the statistical tool used to explain and measure time-to-event data. The term ‘failure’ was used to

define the occurrence of the event of interest even though the event may actually be a 'success' such as recovery from treatment. The term 'survival time' describes the length of time taken for failure to occur. To obtain the survival time, two point times need to be defined. They are time at which an original event occurs to the time of an event of interest.

In most studies in medical field, proportional Cox models are usually used. Cox models only use assumptions such as proportionality of hazard among each variable in a model. In a case where assumptions are violated, Cox models are not suitable to use [10]. This can be determined from the Kaplan-Meier curve. If the curves for each of the independent variables cross each other, then the assumption of proportionality is violated. So, instead of Cox model, parametric survival models will be the best to fit the data. In parametric survival model, it assumes a specific distribution for time variable and produces the best fit model to the data. The distribution of survival times can be described using continuous parametric survival models, by assuming that the hazard has a particular type of shape, with its exact shape determined by one or more parameters which are estimated using the observed data.

Studies by [11] for predicting pregnancy in Friesian Cattle used parametric survival models. The models used were exponential, normal, log-normal, Weibull, logistic and log-logistic distribution. In a colorectal cancer study by [12], parametric survival models were also used to determine the suitable method for the survival of colorectal cancer patients and to identify the independent factors that influenced patient's life expectancy. The parametric models were exponential and Weibull. The distributions used in this study are exponential, log-normal, Weibull and log-logistic. In survival model, T is a non-negative random variable that represents the time until an event of interest occurs.

2.1 Exponential Distribution (Model 1, M1)

A random variable T has the Exponential distribution with the following hazard function h(t), density function f(t) and survival functions S(t):

$$h(t, \lambda) = \lambda \quad (1)$$

$$f(t, \lambda) = \lambda \exp(-\lambda t) \quad (2)$$

$$S(t, \lambda) = \exp(-\lambda t) \quad (3)$$

where t is time until an event of interest occurs and λ is scale parameter.

2.2 Weibull Distribution (Model 2, M2)

As stated by [13], the Weibull distribution can be shown to be a generalization of the exponential distribution with two parametrs and is denoted by T (p,λ). A random variable T has the Weibull distribution with the following hazard function h(t), density function f(t) and survival functions S(t):

$$h(t,\lambda,p) = \lambda p t^{p-1} \qquad p > 0 \text{ (shape)} \quad (4)$$

$$f(t,\lambda,p) = \lambda p t^{p-1} \exp(-\lambda t^{p}) \qquad \lambda > 0 \text{ (scale)} \quad (5)$$

$$S(t,\lambda,p) = \exp(-\lambda\ ) \quad (6)$$

where t is time until an event of interest occurs, p is shape parameter and λ is scale parameter.

2.3 Log-normal Distribution (Model 3, M3)

A random variable T has the Log-normal distribution with the

$$f(t,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma t} e^{-\frac{1}{2\sigma^2}(\log t-\mu)^2} \quad (7)$$

$$S(t,\mu,\sigma^2) = 1 - \Phi\left(\frac{\log t-\mu}{\sigma}\right) \quad (8)$$

$$h(t,\mu,\sigma^2) = f(t)/S(t) \quad (9)$$

where t is time until an event of interest occurs, μ is shape parameter and σ is scale parameter.

2.4 Log-Logistic Distribution (Model 4, M4)

A random variable T has the Log-logistic distribution with the following hazard functions h(t), density function f(t) and survival functions S(t):

$$h(t,\lambda,p) = \frac{\lambda p t^{p-1}}{1+\lambda t^{p}} \quad (10)$$

$$f(t,\lambda,p) = \frac{\lambda p t^{p-1}}{(1+\lambda t^{p})^2} \quad (11)$$

$$S(t,\lambda,p) = \frac{1}{(1+\lambda t^{p})} \quad (12)$$

where t is time until an event of interest occurs, p is shape parameter and λ is scale parameter.

This distribution happens where the hazard rate initially increases and then decreases. At times the distribution can be hump-shaped among the survival time in different types of parametric survival analysis models [14].

In order to select the best parametric model, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and log-likelihood value will be used as criteria for model performance. Akaike Information Criterion (AIC) is a measure of the goodness of fit of a model defined with

parameters estimated by the method of maximum likelihood. The formula of the criterion is:

$$AIC = - 2 \log L + 2 (k) \quad (13)$$

where log L is the log likelihood and k is the degree of freedom. The lowest AIC will be the best fit model.

Bayesian Information Criterion (BIC) is another criterion for performance criteria as AIC is based on the likelihood function. The formula of the criterion is:

$$BIC = - 2 \log L + k \log (n) \quad (14)$$

where log L is the log likelihood, k is the number of freedom and n is the sample size under study. The lowest BIC indicates the best fit model.

A maximum likelihood function can be used to estimate parameters of the parametric models. The log-likelihood function is as follows:

$$\log L (\Theta;y) = \sum_{i=1}^{n} logfi (yi ; \Theta) \quad (15)$$

The highest likelihood value indicates the best fit model.

Thus, the model with lowest AIC value, lowest BIC value and highest log-likelihood value would be the best fit model.

3. RESULTS AND DISCUSSION

Table 1 shows the demographic profile of the patients. From the table, it can be seen that 72 patients suffer from diabetes mellitus, 79 patients suffer from hypertension, 15 patients suffer from anemia, 13 patients suffer from gout, 34 patients are smokers, 58 patients are male and 54 patients have normal BMI.

Table 1: Frequencies for Categorical Variables

Risk Factors	Frequencies (%)
Diabetes Mellitus (Yes)	72 (68.6%)
Hypertension (Yes)	79 (75.2%)
Anemia (Yes)	15 (14.3%)
Gout (Yes)	13 (12.4%)
Smoking (Yes)	34 (32.4%)
Gender (Male)	58 (55%)
BMI (Normal)	54 (51%)

3.1 Model Checking

There are three types of survival analysis. They are standard, mixture and competing risk survival analysis. To determine which type to adopt, model checking has to be performed on the dependent variable using the Kaplan-Meier curve.

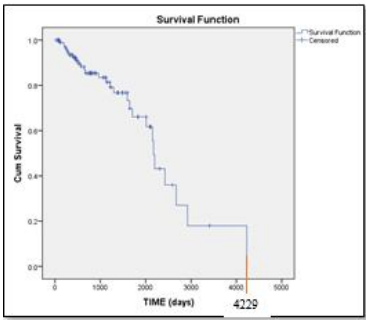
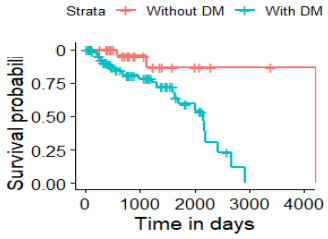
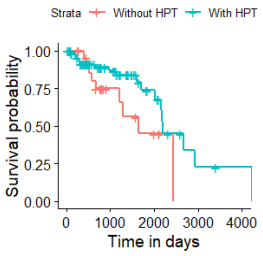


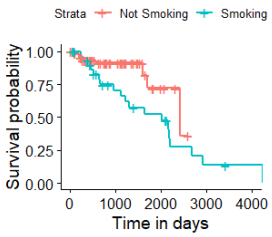
Figure 1: Kaplan-Meier for Survival Time

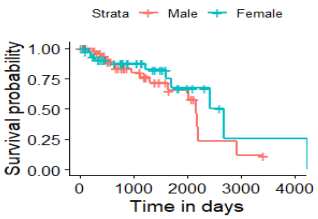
The curve shows that as time increases, the probability of survival decreases. At day 4229, the curve decreases to 0. Because the probability of survival goes to 0, it means the appropriate type of survival analysis to be adopted is standard survival analysis. This decision is supported by the study conducted in [15].

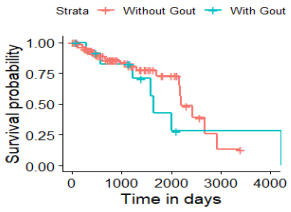
The next step is to determine which survival model to be adopted. There are three survival models, which are parametric, non-parametric and semi-parametric methods. To determine which type of survival model is most appropriate, proportional hazard checking has to be performed on seven independent variables.











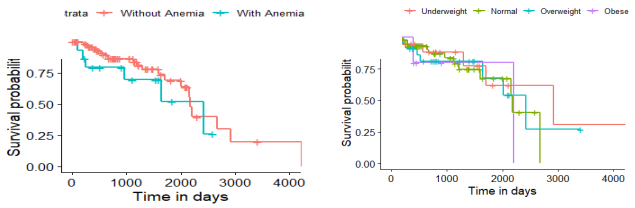


Figure 2: Kaplan-Meier curve for: (a) Diabetes Mellitus (b) Hypertension (c) Smoking (d) Gender (e) Gout (f) Anemia and (g) BMI

Figures 2a-2g show the Kaplan-Meier curve for the independent variables. Figure 2a shows the curves representing patients without diabetes mellitus (without DM) and patients with diabetes mellitus (with DM). It can be seen that the two curves do not cross each other. This indicates that the two survival curves differ significantly. Thus, it can be said that hemodialysis patients without DM have a higher chance of survival than patients with DM.

Figure 2b shows the curve representing patients without hypertension (without HPT) and patients with hypertension (with HPT) crossing each other. This means that the probability of surviving hemodialysis for the two groups of patients is about equal. The same goes to patients who smoke and do not smoke (Figure 2c), male and female (Figure 2d), patients with and without gout (Figure 2e), patients with and without anemia (Figure 2f) and patients who are underweight to obese (Figure 2g). Because there is at least one of the independent variables crossing each other on the Kaplan-Meier curve, the assumption for proportional hazard is violated [10]. Thus, a parametric survival model will be used in this study.

3.2 Model Comparison

All the independent variables/risk factors will be included in all the above distributions. This is called the full model. To determine which variables are significant, independent t-test and chi-square test will be performed. From there, once the significant variables are identified, the variables will be included in another model called the reduced model.

Table 2 shows the results of the tests performed to identify the significant variables for the reduced model.

Table 2: Reduced Model

	P-Value			
Risk Factors	**M1**	**M2**	**M3**	**M4**
(Intercept)	0.000	0.000	0.000	0.000
AGE	0.871	0.792	0.58	0.6611
DM (Yes)	0.051	**0.032**	**0.028**	0.0503
SMOKING (Yes)	**0.015**	0.059	0.062	**0.041**
GOUT (Yes)	0.429	0.515	0.431	0.5984

Table 3 and Table 4 shows the criteria to compare the performance of the full model and reduced model.

Table 3: Comparison of criteria of full model

Criteria	**Full Model**			
	M1	**M2**	**M3**	**M4**
Log-likelihood Value	-237.6	**-227.2**	-229	-228.5
AIC	499.28	**480.30**	482.97	482.97
BIC	499.4	**480.68**	484.28	483.28
Chi-Square	25.39	39.57	41	40.3
Model Sig.	0.008	**0.00042**	0.000024	0.000032
Variable Included	9	9	9	9
Variable Sig.	3	4	5	4

Table 4: Comparison of criteria of reduced model

Criteria	**Reduced Model**			
	M1	**M2**	**M3**	**M4**
Log-likelihood Value	-242.7	**-240.2**	-243.5	-242.6
AIC	495.49	**492.50**	498.99	497.17
BIC	495.51	**492.53**	499.13	497.33
Chi-Square	15.18	13.37	11.91	12.09
Model Sig.	0.0043	**0.0096**	0.018	0.017
Variable Included	4	4	4	4
Variable Sig.	1	1	1	1

It can be seen that in the full model, Weibull distribution has the smallest AIC and BIC and has the highest value of log-likelihood value compared to others. This proved that Weibull distribution is the best model among four parametric models with Chi-square value of 39.57 and the p-value of 0.00042. Hence, the model is significant. Looking at the 'variable significant', Weibull distribution has four significant risk factors that could lead to mortality among hemodialysis patients.

In the reduced model where only the significant variables are included, Weibull distribution also has the smallest AIC and BIC and has the highest value of log-likelihood value compared to others. This shows that Weibull distribution is the best model among four parametric models with

Chi-square value of 13.37 and the p-value of 0.0096. Hence, the model is significant. Looking at the 'variable significant', Weibull distribution has only one risk factor that could lead to mortality among hemodialysis patients.

In both full and reduced model, Weibull distribution is the best model. AIC (480.30) and BIC (480.68) for the full model in Weibull distribution is smaller than AIC (492.50) and BIC (492.53) for the reduced model. However, due to the concept of parsimony, Weibull distribution in reduced model is selected to be the best model. This is because, the number of independent variables used in the reduced model is smaller compared to the number of independent variables used in the full model. Hence, complex model is avoided, because the more variables used in a model, the more complex the model is. Therefore, Weibull distribution in the reduced model will be chosen to identify the risk factors that could lead to mortality among hemodialysis patients. The result shows that diabetes mellitus (0.032) is the only significant risk factor that leads to mortality among hemodialysis patients.

4. CONCLUSION

The study showed diabetes mellitus as the risk factor that influenced mortality among HD patients. The risk of mortality for hemodialysis patients who suffer from diabetes mellitus is 3.9 times higher compared to patients who do not suffer from diabetes mellitus. This result was supported by many previous studies. Study by [3] stated that about 61% of HD patients had diabetes mellitus and diabetes mellitus was found to be the main factor that affect the new number of HD patients in Malaysia. The study by [16] also reported that patients under 50 years old and have diabetes mellitus had higher mortality risk among HD patients compared to patients without diabetes mellitus.

Based on the performance criteria, Weibull distribution was the best fit model in this study among four parametric survival models in predicting risk factors that could lead to mortality among HD patients. This is supported by many previous studies where Weibull distribution was the best model in parametric survival models. A study by [10] showed that Weibull distribution was the best parametric model compared to exponential distribution according to AIC value. A study by [17] in predicting the influential factors among hemodialysis patients also stated that Weibull was the best fit model compared to exponential and log-normal distribution.

ACKNOWLEDGEMENT

We would like to express our gratefulness and appreciation to all dialysis centers that give the permission to collect data used in this research. Appreciation to the University Teknologi MARA for financial support under the Grant Scheme 600-IRMI/MyRA 5/3/BESTARI (039/2017)

REFERENCES

[1] M. do Sameiro-Faria *et al.*, "**Risk factors for mortality in hemodialysis patients: two-year follow-up study**," *Dis. Markers*, vol. 35, no. 6, pp. 791–798, 2013.

[2] H. Yang, Y. H. Chen, T. F. Hsieh, S. Y. Chuang, and M. J. Wu, "**Prediction of Mortality in Incident Hemodialysis Patients: A Validation and Comparison of CHADS2, CHA2DS2, and CCI Scores**," *PLoS One*, vol. 11, no. 5, p. e0154627, 2016.

[3] W.-C. Lee *et al.*, "**The Number of Comorbidities Predicts Renal Outcomes in Patients with Stage 3–5 Chronic Kidney Disease**," *J. Clin. Med.*, vol. 7, no. 12, p. 493, 2018.

[4] J. M. Park *et al.*, "**Survival in patients on hemodialysis: Effect of gender according to body mass index and creatinine**," *PLoS One*, vol. 13, no. 5, pp. 1–14, 2018.

[5] K. C. Abbott *et al.*, "**Body mass index, dialysis modality, and survival: Analysis of the United States Renal Data System Dialysis Morbidity and Mortality Wave II Study**," *Kidney Int.*, vol. 65, no. 2, pp. 597–605, 2004.

[6] S. F. Leavey, K. McCullough, E. Hecking, D. Goodkin, F. K. Port, and E. W. Young, "**Body mass index and mortality in 'healthier' as compared with 'sicker' haemodialysis patients: Results from the dialysis outcomes and practice patterns study (DOPPS)**," *Nephrol. Dial. Transplant.*, vol. 16, no. 12, pp. 2386–2394, 2001.

[7] E. Roddy and M. Doherty, "**Epidemiology of gout**," *Arthritis Res Ther*, vol. 12, p. 223, 2010.

[8] C. Kartsonaki, "**Survival analysis**," *Diagnostic Histopathol.*, vol. 22, no. 7, pp. 263–270, 2016.

[9] D. L. McGee, "**An Introduction to Survival Analysis Using Stata**," *Am. Stat.*, vol. 59, no. 1, pp. 113–113, 2005.

[10] M. Vahedi, M. Mahmoodi, K. Mohammad, S. Ossareh, and H. Zeraati, "**What Is the Best Parametric Survival Models for Analyzing Hemodialysis Data?**," *Glob. J. Health Sci.*, vol. 8, no. 10, pp. 118–126, 2016.

[11] F. D. M. Abdallah and E. A. A. Elfadl, "**Parametric Survival Models for Predicting of Pregnancy in Friesian Cattle**," *Int. J. Stat. Appl.*, vol. 8, no. 3, pp. 129–132, 2018.

[12] M. Yoosefi, A. R. Baghestani, N. Khadembashi, M. Amin, A. A. Baghban, and A. Khosrovirad, "**Survival Analysis of Colorectal Cancer Patients Using Exponentiated Weibull Distribution**," *Int J Cancer Manag.*, vol. 11, no. 3, pp. 2–7, 2018.

[13] V. Vallinayagam, S. Prathap, and P. Venkatesan, "**Parametric Regression Models in the Analysis of Breast Cancer Survival Data**," *Int. J. Sci. Technol.*, vol. 3, no. 3, pp. 163–167, 2014.

[14] A. A. Al-Shomrani, A. I. Shawky, O. H. Arif, and M. Aslam, "**Log-logistic distribution for survival data**

analysis using MCMC," *Springerplus*, vol. 5, no. 1, 2016.

[15] P. K. Swain, G. Grover, and K. Goel, "**Mixture and Non-Mixture Cure Fraction Models Based on Generalized Gompertz Distribution Under Bayesian Approach**," *Tatra Mt. Math. Publ.*, vol. 66, no. 1, pp. 121–135, 2016.

[16] W. H. Lim *et al.*, "**Type 2 diabetes in patients with end-stage kidney disease: influence on cardiovascular disease-related mortality risk**," *Med. J. Aust.*, vol. 209, no. 10, pp. 440–446, 2018.

[17] M. Montaseri, J. Y. Charati, and F. Espahbodi, "**Application of Parametric Models to a Survival Analysis of Hemodialysis Patients**," *Nephrourol Mon.*, vol. 8, no. 6, 2016.