Risk Factors of Lung Cancer Patients: A Survival Analysis with R

Vu-Thanh Nguyen1,2, Thi Diem-Chinh Ho3,4,*

ABSTRACT
Introduction: This paper studies risk factors which can have effects on the survival time of lung cancer patients during the treatment. Methods: The Cox proportional-hazards model has been applied for investigating the association between the survival time of patients and the predictors such as age, gender, the weight of patients, meal, the ECOG, and Karnofsky scores. Results: In the study, we find that the ECOG score, the Karnofsky score evaluated by doctors and the gender are the top three factors that significantly affect the hazard rate. Also, we utilize the estimated model to predict survival probability for the patients. Conclusion: In this article, we intentionally present a complete and detailed guide on how to perform a R-based package in survival analysis step by step as well as how to interpret all output results.

Key words: survival analysis, R software, lung cancer, risk factors, doctors, patients

INTRODUCTION
In scientific research, we sometimes encounter data related to assessing an event over time. For example, a “death” event that describes the time from diagnosis to death or a “relapse” event describes the time from when a given treatment is applied until the recurrence of the disease. Nevertheless, until the end of the study period, not all cases occur “events”. For instance, the study of death caused by lung cancer, from the time of diagnosis, duration of the study was 36 months; not all patients died after the end of the study. Therefore, it can be seen that their time of death is unknown. Hence, this type of research data is characterized as rarely existing in a normal distribution, because some data are complete, some data are censored. Thus, the typical and common method of analysis in this form of data is survival analysis.

Survival analysis is a crucial sector in Statistics to investigate the expected duration of time until one or more events occur. For example, death in biological organisms and failure in mechanical systems. This method is also called reliability theory or reliability analysis in engineering and duration analysis. About this regard, there are many scholars studied it; for instance, Laird and Olivier(1981)1 present the using log-linear analysis techniques to covariance analysis of censored survival data. Hakulinen and Abeywickrama (1985)2 introduced a package to survival analysis. Murray et al. (1993)3 provided a survival analysis of joint replacements. Leggat et al. (1998)4 presented the noncompliance in hemodialysis: predictors and survival analysis. Klein and Moeschberger (2006)5 provided techniques for survival analysis with censored and truncated data.

Survival analysis is a widespread analysis in several medical studies. It is the type of data analysis of an event that occurs at the recorded time, after a certain treatment intervention or after the time the pathology is diagnosed. For example, calculate the risk of death of patients after surgery to remove the liver cancer. Patients will be monitored for the procedure after 2-7 years, statistics of deaths to determine the death rate, the patient’s ability to live at the following during surgery.

In addition, survival analysis can be used to compare the probability of survival of patients after the intervention of two or more certain treatments. Also, it was found that the relationship between the probability of survival of the patient and other factors such as age at intervention, stage of the disease, chemotherapy dose by developing the Cox risk ratio model. Up to the present, the problems of survival analysis are still concerned and researched by several scientists. Readers may refer in Mahmood et al. (2007)6, Strosberg, Gardner, and Kvels, (2009)7, Zhang et al. (2013)8, Stephenson et al. (2015)9, Schlumberger et al. (2017)10, Kyriakopoulos et al. (2018)11, McGregor et al. (2019)12, etc.

The main purposes of this article are to study the association between the survival time of lung patients and the predictors such as age, gender, the weight of patients, meal, the ECOG and Karnofsky scores. Then, we can train a model used for predicting hazard rate and the probability of survival time. Besides, we intentionally present a complete and detailed guide for
survival analysis by using the statistical software R. As we know that, R is a very powerful statistic software, easy to install, and easy to use. Survival analysis plays a tremendously crucial role and immensely profound significant in life. Thus it is exceedingly meaningful to have a study about using the statistical software R in survival analysis.

The paper is organized as follows. The definitions and examples of the ubiquitous functions and the most widespread model in survival analysis are presented in Section 2. The methods used for estimating the ubiquitous functions in survival analysis are also offered in this section. The procedure of applying survival analysis and an application to the real data set is provided in Section 3. Some discussions are illustrated in the next Section 4. Conclusions are stated in the last section.

**METHODOLOGY**

In this section, we first review the most ubiquitous functions for survival analysis, such as survival and hazard function and examples of these functions. We now discuss the survival function in the next section.

**Survival Function**

As we have known, survival time is a non-negative random variable, so if we denote by \( T \) the time of the event, then \( T \geq 0 \). The time considered in the study may be discrete (set of discrete values \( 0 < t_1 < t_2 < t_3 \ldots \) ) or may be continuous \( T \in [0, \infty) \). Let \( S(t) \) be a survival function defined by

\[
S(t) = 1 - F(t) = P(T \geq t) = \begin{cases} \int_0^t f(x) \, dx & \text{as } t \text{ continuous,} \\ \sum_{t_n \geq t} f(t) & \text{as } t \text{ discrete.} \end{cases}
\]

**Hazard Function**

This function is described as the probability that the research object will happen at time \( t \), usually denoted by \( \lambda(t) \) or \( h(t) \). The function \( h(t) \) is defined as follow

\[
h(t) := \lim_{\Delta t \rightarrow 0} \frac{P[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t}
\]

It can be shown that

\[
h(t) = \frac{f(t)}{S(t)}
\]

In fact, we have \( S(t) = 1 - F(t) \) then \( dS(t) = d(1 - F(t)) = -d(F(t)) = -f(t) \, dt \).

Thereafter, we get \( -dS(t) = -f(t) \). Besides, by the definition, we have

\[
h(t) = \lim_{\Delta t \rightarrow 0} \frac{P[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t}
= \lim_{\Delta t \rightarrow 0} \frac{\frac{P[(t \leq T < t + \Delta t)]}{S(t)}}{\Delta t}
= \lim_{\Delta t \rightarrow 0} \frac{1}{S(t)} \frac{\Delta t}{\Delta t}
= \frac{1}{S(t)} \lim_{\Delta t \rightarrow 0} \frac{S(t + \Delta t) - S(t)}{\Delta t}
= - \frac{1}{S(t)} \frac{dS(t)}{dt} = \frac{f(t)}{S(t)}
\]

From this result, we can express

\[-h(t) \, dt = \frac{dS(t)}{S(t)}.
\]

This is equivalent to

\[-h(x) \, dx = \frac{dS(x)}{S(x)}.
\]

Thereafter, we get

\[
\int_0^t -h(x) \, dx = \int_0^t \frac{dS(x)}{S(x)} \, dx
= \log S(t) - \log S(0) = \log S(t),
\]

and

\[
S(t) = \exp \left( \int_0^t -h(x) \, dx \right).
\]

In case of the discrete variables:

\[
h(t_j) = P(T = t_j | T \geq t_j) = \frac{P(T = t_j)}{\sum_{t_k \geq t_j} P(T = t_k)}
= \frac{f(t_j)}{S(t_j)} = \frac{f(t_j)}{\sum_{t_k \geq t_j} f(t_k)}.
\]

The cumulative hazard function is defined by

\[
H(t) = \begin{cases} \int_0^t h(x) \, dx \text{ with continuous variables,} \\ \sum_{t_k < t} h_k \text{ with discrete variables.} \end{cases}
\]

In order for readers to easily access these two functions, we provide two specific examples in the next sub-section.

**Examples of Survival and Hazard Function**

(a) Example 1: Exponential distribution

The exponential density with mean parameter \( \theta \) has the form

\[
f(t) = \frac{1}{\theta} \exp \left( -\frac{t}{\theta} \right).
\]

The survival function is given by

\[
S(t) = 1 - F(t) = \exp \left( -\frac{t}{\theta} \right)
\]
The hazard function is then
\[ h(t) = \frac{f(t)}{S(t)} = \frac{1}{\theta} \exp\left(-\frac{t}{\theta}\right) = \frac{1}{\theta}, \quad \forall t \geq 0. \]

The cumulative hazard is given by
\[ H(t) = \frac{1}{\theta^t}, \quad t \geq 0. \]

(b) Example 2: Weibull distribution
The Weibull density with shape parameter \( \lambda \) and scale parameter \( \theta \) is
\[ f(t) = \frac{\lambda \theta^{-1} t^{\lambda-1}}{\theta^\lambda} \exp\left(-\left(\frac{t}{\theta}\right)^\lambda\right). \]

The survival function can be expressed as follows
\[ S(t) = \int_0^t \frac{\lambda \theta^{-1} u^{\lambda-1}}{\theta^\lambda} \exp\left(-\left(\frac{u}{\theta}\right)^\lambda\right) du \]
\[ = \exp\left[-\left(\frac{t}{\theta}\right)^\lambda\right]. \]

The hazard function is then
\[ h(t) = \frac{f(t)}{S(t)} = \left(\frac{\lambda\theta^{-1}}{\theta^\lambda}\right) \exp\left(-\left(\frac{t}{\theta}\right)^\lambda\right) \]
\[ = \left(\frac{\lambda}{\theta^\lambda}\right) t^{\lambda-1}. \]

The cumulative hazard is given by
\[ H(t) = \left(\frac{1}{\theta^t}\right)^\lambda, \quad t \geq 0. \]

In most analyses, we often consider the appropriate regression model to study the relationship between covariates. It is the same with survival analysis. The Cox regression model is the most widespread regression model in this analysis. Thus we present this model in the next sub-section.

**Cox Regression Model (Cox’s proportional hazards model)**

This model was first proposed by David Cox (1972)\(^\text{13}\). The probability of the endpoint is called the hazard such as death or any other event of interest, e.g., recurrence of the disease. The hazard is modeled as:
\[ h(t) = h_0(t) \exp(\alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n) \quad (4) \]
where \( t \) represents the survival time, \( h(t) \) is the expected hazard at a time \( t \), and \((X_1, X_2, \ldots, X_n)\) is a collection of predictor variables. The coefficient \((\alpha_1, \alpha_2, \ldots, \alpha_n)\) measures the impact of covariates, and \( h_0(t) \) is the baseline hazard at a time \( t \), represents the hazard when all of the predictors (or independent variables) \((X_1, X_2, \ldots, X_n)\) are equal to zero.

The quantities \( \exp(\alpha_i) \) are called hazard ratios (HR). If \( \alpha_i \), or \( HR > 1 \), it indicates that as the value of the \( i \)th covariate increases, the hazard will also be increased. This value is called a bad prognostic factor. In this case, the length of survival decreases. By contrast, if \( \alpha_i < 0 \) it’s called a good prognostic factor. If the hazard ratio equal to 1 corresponding \( \alpha_i = 0 \), the covariate makes no effect.

**Estimation of the survival and hazard function**

We present the two most widespread methods to estimate the survival and hazard function that is Kaplan-Meier estimator and Nelson-Aalen estimator. Let \( t_i \) be a time when at least one event happened. Suppose that \( t_i \) is arranged in order: \( 0 < t_1 < t_2 < \ldots < t_i, \) \( d_i \) is the number of events (e.g., deaths) that happened at the time \( t_i, \) \( n_i \) is the number of objects that occur at \( t_i \) or later, and let \( r_i \) be the number of individuals at risk (i.e., alive and not censored) just before to time \( t_i \).

**Kaplan-Meier estimator**

This estimator is a non-parametric maximum estimator for the survival function \( S_i \). This method is proposed by Kaplan and Meier (1958)\(^\text{14}\). It can be expressed as follows
\[ \hat{S}(t) = \prod_{t_i \leq t} (1 - \hat{h}_i) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \quad (5) \]

**Nelson-Aalen estimator**

The Nelson-Aalen estimator is a non-parametric estimator that can be utilized to estimate the hazard function. Because there is no need for distribution assumptions, a crucial use of the estimator is to test the appropriateness of parameter models graphically, and this is exactly why this method was first introduced by Nelson (1969, 1972)\(^\text{15,16}\). Independent from Nelson, Altshuler (1970)\(^\text{17}\) also studied this issue. Later, Aalen (1978)\(^\text{18}\) expanded its use beyond survival data and established competitive risk, and studied its small and large sample properties by martingale. Nowadays, this estimator is called the Nelson-Aalen estimator. This function can be expressed as follows
\[ \hat{H}(t) = \sum_{t_i < t} \frac{d_i}{r_i} = \sum_{t_i < t} \hat{h}_i \quad (6) \]

**DATA AND ANALYSIS RESULTS**

The procedure of using R in survival analysis

In this section, we introduce the procedure of survival analysis using R language as well as the usage of some functions.
Step 1. Determining variables related to time factors such as lifetime, survival time, and failure time. Before using the R-system software for survival analysis, we first have to load the survival package into the working environment by using the command library(survival).

Step 2. Calculating descriptive statistics, plot some graphs for the data set, for instance, using plot, Survand survfit built in the survival package.

Step 3. Applying the Cox regression model to find out the risk factors which have effects to the hazard risk \( h(u) \). The command is coxph().

The simple usage of the function coxph() is given below: 

```r
coxph(combined variables ~ independent variables, data)
```

or it can be written as follows

```r
coxph(Surv(variable 1, variable 2) ~ independent variables, data)
```

Step 4. We can use cox.zph() to test the hypothesis of risk factors: 

```r
cox.zph(coxph(Surv(variable 1, variable 2) ~ independent variables, data))
```

In order to help readers that can easily use the statistical software R in survival analysis in practice, we introduce analysis with real data in the next section.

**Application and analysis results**

Firstly, we load the survival package to be developed by Terry Therneau et al. into the working environment:

```r
> library(survival)
```

In this section, we do analysis on the lung cancer data set, which is taken from the North Central Cancer Treatment Group. This data set is named "lung". The data set has 228 rows and 10 columns, with 10 variables described as follows:

1. inst: Hospital and organization code.
2. time: Survival time in days.
3. status: censored: = 1, dead: = 2.
4. age: age of patient.
5. sex: Men = 1, Women = 2.
6. ph.ecog: Score for ECOG effect (good = 0, death = 5).
7. ph.karno: The scores for Karnofsky’s effects are evaluated by doctors.
8. pat.karno: The scores for the effect of Karnofsky was assessed by patients.
9. meal.cal: The number of calories consumed during the meal.
10. wt.loss: Weight loss for the last 6 months.

The ECOG scale first appeared in the medical literature in 1960. It describes a patient’s level of functioning in terms of their ability, between 0 and 5, to care for themselves, daily activity, and physical ability, (walking, working, etc.). The table below was developed by the Eastern Cooperative Oncology Group, Robert L. Comis, MD, Group Chair, based on the paper of Oken M, Creech R, Tormey D, et al. (1982).

Karnofsky index (in Karnofsky D, Burchenal J., 1994) similar to ECOG, between 100 and 0, was introduced in a textbook in 1949. The table below displays Karnofsky index.

Using the data above, we focus on how to estimate the probability of survival time for patients and to find out risk factors which can cause deaths to the patients based on the Cox regression model. Next, we can take a look on the first 6 rows and last 6 rows of the data set by using the command head(lung) and tail(lung). After executing these commands, the result is depicted in Figure 1.

Note that NA is missing values. We now can compute descriptive statistics by installing and loading some packages, as shown below:

```r
> install.packages("fBasics")
> install.packages("tseries")
> library(fBasics)
> library(tseries)
```

Because our data set is named "lung", so we execute the command:

```r
> basicStats(lung)
```

For simplicity, we can utilize the round(result, 2) to get 2 decimal places. The general of the round function in R is round(result, n) to get \( n \) decimal places.

```r
> round(basicStats(lung),2)
```

After performing the above command, the result of the descriptive statistics for the data set “lung” is depicted in Figure 2.

As can be seen in Figure 2, the average lifetime of the patients is more than 305 days. The average of their age is relatively high (62.45). The number of men accounts for a slightly higher proportion (1.39). The score for ECOG is less than 1, while Karnofsky is quite stable rated by both doctors and patients (81.94 and 79.96). The average calories consumed by the study subjects is 928.78.

Regarding the standard deviation (sd), meal.cal attains the largest sd of 402.17, which is almost double the standard deviation of the second-highest quantity of 201.65 for the survival time. Age with a deviation is no more than 10 (9.07), and weight loss and scores of the doctor’s and patients Karnofsky have sd...
Table 1: ECOG performance status Grade

| GRADE | STATUS |
|-------|--------|
| 0     | Fully active, able to carry on all pre-disease performance without restriction. |
| 1     | Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light housework. |
| 2     | Ambulatory and capable of all self-care but unable to carry out any work activities; up and about more than 50% of waking hours. |
| 3     | Capable of only limited self-care; confined to bed or chair more than 50% of waking hours. |
| 4     | Completely disabled; cannot carry on any self-care; totally confined to bed or chair. |
| 5     | Dead |

Table 2: Karnofsky performance status index

| Index | STATUS |
|-------|--------|
| 100   | Normal, no complaints; no evidence of disease |
| 90    | Able to carry on normal activity; minor signs or symptoms of the disease |
| 80    | Normal activity with effort, some signs or symptoms of the disease |
| 70    | Cares for self but unable to carry on normal activity or to do active work |
| 60    | Requires occasional assistance but can care for most of the personal needs |
| 50    | Requires considerable assistance and frequent medical care |
| 40    | Disabled; requires special care and assistance |
| 30    | Severely disabled; hospitalization is indicated although death not imminent |
| 20    | Very ill; hospitalization and active, supportive care necessary |
| 10    | Moribund |
| 0     | Dead |

Figure 1: All variables used in the lung dataset, which is taken from the North Central
in the range from 12 to 15. Meanwhile, the deviation of ECOG is not more than 1. Among variables, the median of meal.cal is still the largest value of 975 and follow by the number of days the patients lived having median of 255.5. The Karnofsky score in both types is 80. Median of the status is 2 indicating that the number of subjects censored is less than the number of deaths. Median of sex is 1 showing that there are more men over women in the study. Median values of the age, the weight loss after 6 months, and the ECOG score is 63 years, 7 kg and 1, respectively.

The maximum and minimum values (max, min) of survival time are 1022 and 5, respectively. The age of participants in the study is from 39 to 82; the score for ECOG ranges from 0 to 3. The highest score for both Karnofsky types is 100, and we see a difference in the minimum values (50 for doctors score, 30 for patients score). Calories per meal have been increased from 96 to 2600 and after 6 months, some patients lost up to 68 kg in weight, while the other patients increased the weight up to 24 kg.

The standard error (SE) of calories and age is quite large (29.89 and 13.95), while SE values of the rest variables are smaller than 1.

In the survival package, there are Survand survfit functions that give an overview of the data set. The Surv function is used to create a compound variable, for instance, a combination of time and status as in the following command:

```r
> time=lung[,2]
> status=lung[,3]
> survival.time = Surv(time, status==2)
```

After executing the above command, the result is depicted in Figure 3.

Note that this variable is only valuable and meaningful for the analysis in R, but in reality maybe we do not need it.

For the survfit function, it is also quite simple, we only need to provide two parameters: time and status as in the following example:

```r
> survfit(Surv(time,status==2) ~ 1)
```

If there is already an object “survtime”, then we simply call the command:

```r
> survfit(survival.time ~ 1)
```

After executing the above command, it outputs the result as follows:

```r
Call: survfit(formula = survival.time ~ 1)
N events median 0.95LCL 0.95UCL
228 165 310 285 363
```

In order to illustrate survival probability, we can represent it through a graph. We execute the following command:

```r
> plot(survfit(survival.time ~ 1),
xlab="Time",ylab="Cumulative survival probability", col=c("red", "black","black"))
```

After performing the above command, the result in R is provided as in Figure 4.

After we have thoroughly investigated the data set, we want to study how the factors affect the hazard rate h(t). The Cox regression model helps us address this problem. We can estimate this model by the function coxph(formula, data).
Figure 3: Life time of the patients. The plus sign “+” indicates that the patients were censored or alive, otherwise they passed away during the study.

Figure 4: Survival probability of the patients with 95% confidence interval. The time counted in days and the cumulative survival probability of $S(t)$ of the objects. The central red line is the cumulative probability of $S(t)$, and two black lines are 95% confidence intervals. The 95% confidence interval of this example is relatively good, and the interval is changing from narrow and larger width, which can help us to realize the change of the survival probability of patients over time.
We first use the Cox regression model to check relationship between the compound variable (time and status) and all remaining independent variables by the following command:

```
> coxph(Surv(time,status) ~ age+sex+ph.ecog +ph.karno+pat.karno+meal.cal+wt.loss, data=lung)
```

After performing the above command, we have the results presented as follows.

```
> pkh=coxph(Surv(time,status) ~ age+sex+ph.ecog +ph.karno+pat.karno+meal.cal+wt.loss, data=lung)
> pkh
Call: coxph(formula = Surv(time, status) ~ age + sex + ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss, data = lung)

coef exp(coef) se(coef) z p-value
age 1.065e-02 1.0107 0.9894 1.0004 0.000192
sex -5.509e-01 0.5765 0.3889 0.89791
ph.ecog 7.342e-01 2.0838 1.3452 3.2277
ph.karno 2.246e-02 9.8777 1.0125 0.9722 1.0034
pat.karno -1.242e-02 9.858e-01 8.0000 9.9995 1.0005
meal.cal 3.329e-05 1.0000 1.0000 1.0000 1.0000
wt.loss -1.433e-02 9.8777 1.0125 0.9722 1.0034
Likelihood ratio test = 28.33 on 7 df, p=2e-04
n= 168, number of events= 121
```

It can be seen that, this is the results based on the influence of age, sex, Ecog and Karnofsky scores of doctors and patients, calories per meal and weight lost over 6 months. Results show that the variables sex, ph.ecog, and ph.karno are significant since p-values are less than 0.05. However, the covariates of age, pat.karno, meal.call, and wt.loss are failed to be significant.

From this result, the p-values for Likelihood ratio test, Wald test and Score (logrank) test indicate that the model is significant. These tests are applied to evaluate the null hypothesis that all of coefficient $a$ of model is zero. So in this case, the null hypothesis is rejected. We can see that the covariates sex, ph.ecog, and ph.karno are significant since p-values are less than 0.05. However, the covariates of age, pat.karno, meal.call, and wt.loss are failed to be significant.

For the purpose of predicting the survival probability of an event, we split data into training data set and testing data set with the ratios 80% and 20% of the whole data, respectively. With the training data set,
we run the Cox survival model to learn the coefficients. Then, we use the testing data set to predict the survival probability of the event.

```r
> set.seed(300)
> training.samples=createDataPartition(lung$time,p=0.8, list=FALSE)
> training.data<-lung[training.samples,]
> test.data<- lung[-training.samples,]

pkh=coxph(Surv(time,status)~age+sex+ph.ecog+ph.karno+pat.karno+meal.cal+wt.loss, data=training.data)
> test=test.data[-c(1,2,3)]
> test1 = test.data[-1]
> predict(pkh,newdata=test1, "expected")
> exp(-predict(pkh,newdata=test1, "expected"))

[1] 0.57250330 NA NA NA 0.93834019
[6] 0.60755065 NA 0.96866930 NA 0.17090105
[11] 0.83522433 0.35415900 0.74692248 0.86582180
0.03103017
[16] 0.58304061 0.17535892 0.86821706 0.90241398
[21] 0.60967940 NA 0.75159416 0.35146879
0.37332993
[26] 0.86343272 0.43038027 NA 0.31838746 NA
[31] 0.67358530 0.61329536 0.64212740 NA 0.81973135
[36] NA NA 0.77227962 0.95996078 NA
[41] 0.80716582 NA 0.72431187 0.84400730
Dataframe of test1 with the first six rows:
> head(test1)
  time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
 4 210 2 57 1 90 60 1150 11
12 654 2 68 2 70 70 NA 23
13 728 2 68 2 90 90 NA 5
14 71 2 60 1 NA 60 70 1225 32
22 81 2 49 2 100 70 1175 -8
24 371 2 58 1 90 100 975 13
At the first row in test1 data set, we can state that the probability that a 57-year-old man is still alive after 210 days is 0.5725. Also, the model can't define probability for some events because of missing data.

DISCUSSIONS

The paper presents a completed and detailed guide survival analysis and its application by using the statistical software R. Survival analysis plays a tremendously crucial role an immensely profound significant in life. Thus, it is exceedingly meaningful to have a study about using the statistical software R in survival analysis.

This paper is also a valuable reference for faculty members as well as students in probability and statistics. Besides, it can also help those who are interested in this area. The article is very detailed and complete about survival analysis, so it is easy for people to access and understand it.

In general, it can be seen that the functions in R used for survival analysis will work well if our data set does not contain missing values. In practice, however, we often encounter data sets that contain missing values. Therefore, we cannot immediately apply the available functions in R for survival analysis. Usually, the effective solution for this situation is to combine with methods used for solving the missing data problems. Reranging missing data problems, they have also been studied and mentioned by different researchers. For instance, Pho et al. (2019b)\(^2\) reviewed three update methods to solve the issues with missing data. Besides, we refer to Little (1992)\(^2\), Horton and Kleinman (2006)\(^5\), Mahmoudi et al. (2020a)\(^24\), etc; for further details. This will also be a potential research direction if we can combine the methods used for dealing with missing values and the methods of survival analysis.

In this study, we capture the relationship between survival time and risk factors of lung cancer patients by only the Cox regression. For future work, it will be a more powerful tool if we can utilize different machine learning models such as survival tree, random forest, etc. Also, we can think of applying cause-effect relationships inference in survival analysis. Such relationships can be modeled by the Directed Acyclic Graph (DAG), and the effect scores of the risk factors can be determined by using the Treenet models, see, e.g., Changpetch (2016)\(^25\). Time series analysis and survival analysis lead to future object changes, so understanding and working together will yield more reliable results. These research tools have an important practical application, such as the study of the Covid-19 disease. About related research topics can look in Maleki et al.\(^26,27\), Pho, K. H.\(^28\).

CONCLUSIONS

In this study, we have presented the most ubiquitous functions in survival analysis, consisting of survival and hazard function and examples of these functions. In addition, we also reviewed the Cox regression model which is one of the most widespread model in survival analysis. Moreover, we provided the approach to estimating the survival and hazard function as well as the procedure of using R in survival analysis. As an application, the data set of lung cancer has been used to analyze risk factors which can cause
deaths of the patients. Furthermore, we have also introduced some extensive research directions such as doing survival analysis with missing data and the study of the cause-effect relationships in survival analysis.

REFERENCES

1. Laird N, Olivier D. Covariance analysis of censored survival data using log-linear analysis techniques. Journal of the American Statistical Association. 1981;76(374):231–240. Available from: https://doi.org/10.1080/01621459.1981.10477634.

2. Hakulinen T, Abeywickrama KH. A computer program package for relative survival analysis. Computer programs in biomedicine. 1985;19(2-3):197–207. Available from: https://doi.org/10.1016/0160-482X(85)90011-X.

3. Murray D, Carr A, Bulpin D. Survival analysis of joint replacements. The Journal of bone and joint surgery British. 1993;75(5):697–704. Available from: https://doi.org/10.1302/0301-620X.75B5.8376423.

4. Leggat J, Orzol SM, Hulbert-Shearon TE, Golper TA, Jones CA, Held PJ, et al. Noncompliance in hemodialysis: predictors and survival analysis. American Journal of Kidney Diseases. 1998;32(1):139–145. PMID: 9669435. Available from: https://doi.org/10.1053/ajkd.1998.v32.pm9669435.

5. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. Springer Science & Business Media.

6. Mahmood A, Raymond GV, Dubey P, Peters C, Moser HW. Survival analysis of haematopoietic cell transplantation for childhood cerebral x-linked adrenoleukodystrophy: A comparison study. The Lancet Neurology. 2007;6(8):687–692. Available from: https://doi.org/10.1016/S1474-4422(07)70177-1.

7. Strosberg J, Gardner N, Kvols L. Survival and prognostic factor analysis in patients with metastatic pancreatic endocrine carcinomas. Pancreas. 2009;38(3):255–258. PMID: 19066493. Available from: https://doi.org/10.1097/01.TP.00003181917e46.

8. Zhang W, Ota T, Shihidrar V, Chien J, Wu B, Kung R. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. Plos Comput Biol. 2013;9(3):e1002975. PMID: 23555212. Available from: https://doi.org/10.1371/journal.pcbi.1002975.

9. Stephenson AL, Tom M, Berthiaume Y, Singer LG, Aaron SD, Whitmore G, et al. A contemporary survival analysis of indeterminate factor analysis in patients with metastatic pancreatic adenocarcinoma. The Lancet. 2004;363(9409):995–999. Available from: https://doi.org/10.1016/S0140-6736(04)16741-9.

10. Slumberger M, Elsise R, uiller SM, sfsi PS, Brose M, Shah M, et al. Overall survival analysis of exam, a phase 20 iii trial of cabozantinib in patients with radiographically progressive medullary thyroid carcinoma. Annals of Oncology. 2017;28(11):2813–2819. PMID: 29045520. Available from: https://doi.org/10.1093/annonc/mdx479.

11. Kyriakopoulos CE, Chen YH, Carducci MA, et al. Chemohormonal therapy in metastatic orrme-sensitive prostate cancer: long-term survival analysis of the randomized phase iii e3805 charted trial. Journal of Clinical Oncology. 2018;36(11):1080. PMID: 29384722. Available from: https://doi.org/10.1200/JCO.2017.75.3657.

12. McGregor D, Palarea-Albaladejo J, Dall P, Hron K, Chastin S. Cox regression survival analysis with compositional covariates: Application to modelling mortality risk from 24- physical activity patterns. Statistical methods in medical research. 2020;29(5):1447–1465. PMID: 31342855. Available from: https://doi.org/10.1177/0962280219864125.

13. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187–202. Available from: https://doi.org/10.1111/j.2517-6166.1972.tb00899.x.

14. Kaplan EL, Meier P. Non-parametric estimation from incomplete observations. Journal of the American statistical association. 1958;53(282):457–481. Available from: https://doi.org/10.1080/01621459.1958.10501452.

15. Nelson W. Hazard plotting for incomplete failure data. Journal of Quality Technology. 1969;1(1):27–52. Available from: https://doi.org/10.1080/00224065.1969.11980344.

16. Nelson W. Theory and applications of hazard plotting for censored failure data. Technometrics. 1972;14(4):945–966. Available from: https://doi.org/10.1080/00401706.1972.10488991.

17. Altshuler B. Theory for the measurement of competing risks in animal experiments. Mathematical Biosciences. 1976;1(1):11. Available from: https://doi.org/10.1016/0025-5564(76)90052-0.

18. Aalen O. Non-parametric inference for a family of counting processes. The Annals of Statistics. 1978;6:701–726. Available from: https://doi.org/10.1214/aos/1176344247.

19. Therneau TM, Lumley T. Package survival. 2014;.

20. Oken MM, Creech RH, Torrey DC, Horton J, Davis TE, McFadden ET, et al. Toxicity and response criteria of the eastern cooperative oncology group. American journal of clinical oncology. 1982;5(6):649–656. PMID: 7165009. Available from: https://doi.org/10.1097/00000421-198212000-00014.

21. Karnofsky D, Burchenal J. The clinical evaluation of chemotherapy agents in cancer. CM Mac Leod (A–D), e Valuation of Chemotherapeutic Agents. 1949;p. 191.

22. Pho KH, Ly S, Ly S, Lukusa TM. Comparison among akaike information criterion, Bayesian information criterion and Vuong’s test in model selection: A case study of violated speed regulation in Taiwan. Journal of Advanced Engineering and Computation. 2019;3(1):293–303. Available from: https://doi.org/10.25073/ijace.201931.220.

23. Little RJ. Regression with missing x’s: a review. Journal of the American statistical association. 1992;87(420):1227–1237. Available from: https://doi.org/10.1080/01621459.1992.10476282.

24. Mahmoudi MR, Heydari MH, Avazadeh Z, Pho KH. Goodness of fit test for almost cyclostationary processes. Digital Signal Processing. 2020;66:102597. Available from: https://doi.org/10.1016/j.dsp.2019.102597.

25. Changpetch P, Haughton D, Le M, Ly S, Nguyen P, Thach T. Alcohol consumption in Thailand: A study of the associations between Alcohol, Tobacco, Gambling, and Demographic Factors. Tobacco, Gambling, and Demographic Factors. 2016;Available from: https://doi.org/10.2139/ssrn.2732844.

26. Maleki M, Mahmoudi MR, Wraith D, Pho KH. Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel Medicine and Infectious Disease. 2020.p 101742. PMID: 32405266. Available from: https://doi.org/10.1016/j.tmaid.2020.101742.

27. Maleki M, Mahmoudi MR, Heydari MH, Pho KH. Modeling and Forecasting the Spread and Death Rate of Coronavirus (COVID-19) in the World using Time Series Models. Chaos, Solitons & Fractals. 2020.p 110151. Available from: https://doi.org/10.1016/j.chaos.2020.110151.

28. Pho KH, Heydari MH, Tuan BA, Mahmoudi MR. Numerical study of nonlinear 2D optimal control problems with multi-term variable-order fractional derivatives in the Atangana-Baleanu-Caputo sense. Chaos, Solitons & Fractals. 2020;134:109695. Available from: https://doi.org/10.1016/j.chaos.2020.109695.