ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 1609-1617 (2020)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

Time-to-Event Analysis for Recovery from Coronavirus Disease (COVID-19): A Case Study on Wuhan and Elsewhere in China from Jan 1 to Feb 11, 2020

Murtada Khalafallah Elbashir*,[1,2], Saleh N. Almuayqil[1]

[1] *Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakaka, 72311, Kingdom of Saudi Arabia*

[2] *Department of Computer Science, Faculty of Mathematical and Computer Sciences, University of Gezira, Madani, 585-0032, Sudan*

ARTICLE INFO

Article history:
Received: 19 September, 2020
Accepted: 21 December, 2020
Online: 25 December, 2020

Keywords:
Kaplan-Meier method
Cox regression model
Proportional Hazard model

ABSTRACT

COVID-19 is a viral disease that became a pandemic representing a very great challenge worldwide. The purpose of this article is to analyze COVID-19 patients' data based on time-to-event analysis and identify the factors that affect the recovery time from COVID-19. The datasets that are used in this study are for cases that are clinically diagnosed and confirmed where the date of onset is recorded in Wuhan and elsewhere in China from Jan 1 to Feb 11, 2020. We used the regression imputation technique to replace the missing dates in the onset-symptoms based on the dates of the report. We fitted the Kaplan-Meier estimator and Cox regression model to our data. The predictor variables (factors) that we used are age, sex, and onset time to hospitalization. The results show that the young age group is better than the old age group in recovering from COVID-19 (the p-value of the log-rank is 0.00012) and at any time 1.9 as many patients in the young age group are having an event (recovery) proportionally to the old age group. Also, the results show that there is a non-significant difference between male and female groups in recovering from COVID-19 (the p-value of the log-rank is 0.63). The results also show that the early time to hospitalization group can recover from COVID-19 better than the late time to hospitalization group (the p-value of the log-rank is 0.0052). This study demonstrates the association of recovery time from COVID-19 with age, sex, and time to hospitalization.

1 Introduction

The novel coronavirus disease also known as COVID-19 is a viral disease that became a pandemic and turn out to be a great challenge that the world faced since world war two. This virus is originated in Wuhan city, which located in the Hubei province of China and it is spreading at a fast rate around the globe. The diseases caused by viral infection continue to emerge and raise a serious issue in public health worldwide. Several viral epidemics appeared in the last twenty years [1]. In 2002 the severe acute respiratory syndrome coronavirus, which is known as SARS-CoV, which is still circulating in China [2]–[4] has appeared followed by H1N1 influenza in 2009. Most recently in 2012, the Middle East respiratory syndrome coronavirus, which is known as MERS-CoV have been recorded. The main component of the viral genome is a positive-stranded RNA and it has a different structure [5]. There are four genera of Coronavirinae family: α, β, γ, and δ. it is believed that there is a viral gene in wild animals since it has been isolated from bats and other animals [6]. The novel COVID-19 causes mild to moderate respiratory illness, but some people and people with health problems can develop serious illness. Worldwide this disease affected more than five billion people and the number of people who died due to the infection with it exceeds five hundred thousand according to the World Health Organization (WHO) report on the time of writing this research. According to WHO the mild or asymptomatic COVID-19 infections represent 80% of the cases while the severe infections, which require oxygen and critical infections, which require ventilation represents 15% and 5% of the cases respectively. So far, the mortality for COVID-19, which is the total number of deaths divided by the total cases is 5% (this percentage is calculated according to figures that are taken from the WHO web site on 09 July 2020, which shows a total infection of (12,196,982) and total deaths of (552,781)). This mortality is higher than that of seasonal influenza, which is below 1% according to WHO.

*Corresponding Author: Murtada K. Elbashir, Jouf University, mkelfaki@ju.edu.sa

https://dx.doi.org/10.25046/aj0506192

Normally, data mining and machine learning methods can be used to analyze datasets of biomedical data [7]. When data include survival data, it requires a different analysis approach. This approach or the study of the time from the entry of a study until a subsequent event occurs is known as survival analysis. Survival analysis is applied in different disciplines such as medicine, engineering, social sciences or behavioral sciences and biology [8]–[16]. When it is applied to medicine, survival analysis is used to study people at risk of experiencing a negative event such as death, where the name survival analysis comes from. Survival analysis is also applicable to areas other than mortality such as analyzing the time taken to recover from certain diseases or the time taken to practice certain exercises to maximum tolerance [17]–[20]. Normally we compare two or more groups of patients with respect to the time of event. More than one event can be considered in the same analysis, but we normally take one event at a time as the event of interest in the study and it can be death or recovery [21].

Many methods can be used for survival analysis, these methods include Kaplan-Meier method which is an estimator of survival probabilities [19, 22] and the Cox regression model, which is now known as the Cox Proportional Hazard Model (CPHM) [23]. These two methods are considered among the methods that contribute significantly to the development of the survival analysis field.

Many studies were conducted to model the survival time and to predict the mortality risk for COVID-19. Guillermo Salinas-Escudero et al. applied survival analysis to study the effect of COVID-19 in the Mexican [24]. The factors they used include age, sex, comorbidities, hospitalization, and admission to the intensive care unit. They applied the Kaplan-Meier and Cox regression models to their data. Their results show that men and older people have higher mortality than women and young people respectively. Monira Mollazehi et al. modeled survival time to recover from COVID-19 [25]. They used data from Singapore in the period between January 23 and March 13, 2020. Their purpose is to identify the factors affecting the recovery time from COVID-19. They used patient's age and nationality as predictors and they found that younger patients can recover from COVID-19 faster than old patients and Singaporean patients can recover faster than non- Singaporean. They compared the results of different models and they found that the Weibull model is the best in fitting their data. Using the Weibull model, they obtained a Hazard rate of 1.01 and 0.76 for age and nationality respectively. Qinxia Wang et al. used survival-convolution models to model the duration of the patient remaining infectious to others [26]. Noam Barda et al proposed a hybrid methodology to construct a multivariable prediction model. In their hybrid method, they used a baseline model which they trained on population data to discriminate the risk then they used a multicalibration algorithm for the risk predation [27]

Different factors may have an influence on the mortality or the recovery time from COVID-19. These factors can be used to divide the patients into two or more groups and they include age, gender, and time from acquiring the illness to hospitalization. This study aims to investigate whether these factors affect recovery time. The datasets that are used in this study are downloaded from Github (`https://github.com/mrc-ide/COVID19_CFR_submission`). From these datasets, we used two datasets. The first one is for cases that died from COVID-19 in Hubei and the second dataset is for patients returning to their home, which obtained from six flights that departed between Jan 30 and Feb 1, 2020.

The rest of the paper is organized as follows: The next section describes the materials and methods. The material and methods section starts by showing how we prepared the dataset that we used followed by describing the imputation technique we used to replace the missing data. Kaplan–Meier survival curve, Log-rank test, and Cox proportional hazards (PH) model also are explained in the materials and methods section. we present the results and the discussion in the third section and the conclusion in the last section.

2 Material and Methods

2.1 *Dataset*

The datasets that we used in this study are for cases that are clinically diagnosed and confirmed where the date of onset is recorded in Wuhan and elsewhere in China from Jan 1 to Feb 11, 2020. From these datasets, we used two datasets. The first one is for cases that died from COVID-19 in Hubei. It contains the features: sex, age, date of symptom onset, date of hospitalization, and date of death or recovery from COVID-19. Some of the data for the date of symptom onset are not available for some cases so used imputation based on regression to replace the missing data. The second dataset is for patients returning to their home, which obtained from six flights that departed between Jan 30 and Feb 1, 2020. Also, the cases with incomplete date of symptom onset were replaced using regression imputation and then we merged the two datasets. We removed the cases where the sex or the age or date of hospitalization are not available and we end up with 693 cases, which represent recovered and died patients. The used datasets were downloaded from Github (`https://github.com/mrc-ide/COVID19_CFR_submission`). They were used by [28], which extracted it from WHO–China Joint Mission report to estimates the severity of COVID-19 based on the model-based analysis.

2.2 *Regression imputation*

The datasets that we downloaded has missing data on the date onset symptoms, therefore, instead of deleting all the cases that have missing data, we need imputation to replace these missing data with estimated values, because it is important to have the timing of the onset-symptoms to study the recovery time. We used regression imputation to preserves all cases by replacing the missing date onset symptoms with a probable value estimated by the date of the report because it is clear from Figure 1 that there is a strong correlation between these two dates. In Figure 2, the scatter plot shows the relationship between these two dates, and the value of R^2 (0.7266) emphasizes the strength of this relationship. The model that is used to estimate the missing data in the date onset symptoms is $y = 0.8227x - 0.5428$ also shown in Figure 2, where x and y represent the report date and the onset symptoms respectively. Preserving the cases with missing data using regression imputation has several advantages. In addition to avoiding the deletion of the cases with missing data that can alter the variance of the shape distribution, it can also substitute the missing value based on another variable and no novel information will be added therefore we will be having an increased sample size and therefore a reduced standard error [29]–[31].

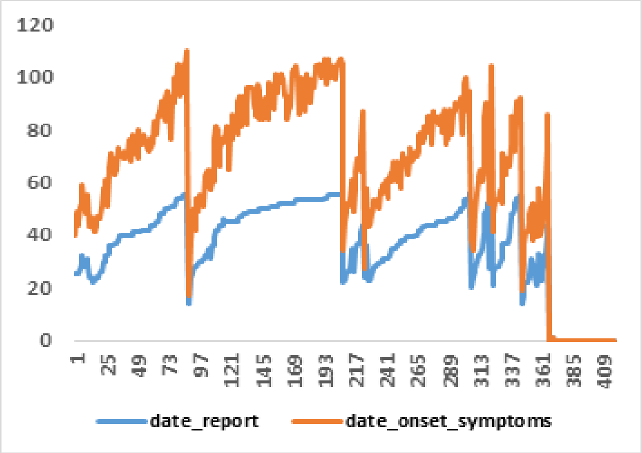

Figure 1: The correlation between date_report and date_onset_symptoms.

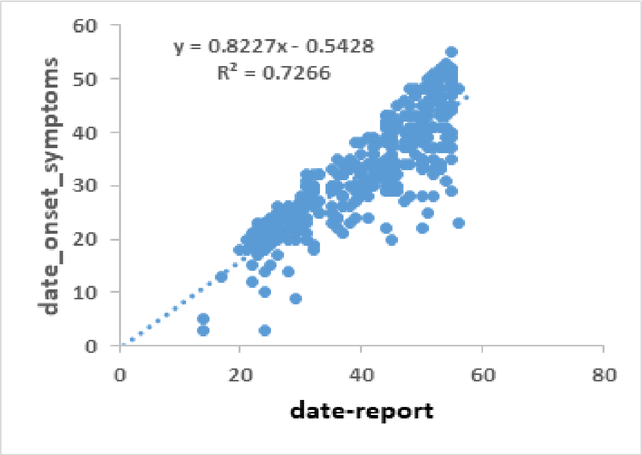

Figure 2: Regression prediction for date_onset_symptoms.

2.3 Kaplan–Meier survival curve

In Kaplan–Meier survival curve, the survival times that include the censored data (the observation that does not get the event) is assumed to be $t_1, t_2, ..., t_n$. These times are entered to the study ordered by increasing duration of a group of n subjects, We can estimate the proportion (survival rate) of subjects $S(t)$ surviving beyond any follow-up time t_p as [17]:

$$S(t) = \frac{(r_1 - d_1)}{r_1} \times \frac{(r_2 - d_2)}{r_2} ... \times ... \frac{(r_p - d_p)}{r_p} \tag{1}$$

Here r_i represents the number of subjects alive just before time t_i given that t_p is the largest survival time and i is any value between 1 and p, d_i represents the number of subjects who died at the time t_i, therefore $d_i = 0$ for censored observations. Before the occurrence of the first event all the patients are alive, therefore, $S(t) = 1$. Considering time t_i, where the number of events(deaths) is d_i and the number of alive is r_i just before t_i then $S(t_i)$ can be calculated as:

$$S(t_i) = \frac{(r_i - d_i)}{r_i} \times S(t_{i-1}) \tag{2}$$

In the censored data we will not have information about the survival time, therefore, $S(t_i)$ will not be calculated for censored observations since the survival curve will not change at the time of a censored observation. At the next event, the number of patients at risk is reduced by the number of censored observations between the two events [32].

2.4 Log-rank test

Normally, we need to compare two survival curves of two groups. For this sake, we use Log-rank test, which is related to a test that uses the logarithms of the ranks of the data and it is used under the assumptions: i) the survival times are continuous or ordinal, ii) one group's risk of an event relative to the other does not change with time. When the death event occurs at time t_i then we will consider the total number alive (r_i) and the total number still alive up to the time t_i in a specific group (say group A) r_{Ai}. Consider that d_i is the total number of deaths i.e event at the time t_i. Then the expected number of deaths in group A at time t_i can be calculated as

$$E_{Ai} = \frac{r_{Ai}}{r_i} \times d_i \tag{3}$$

Then the total number of expected deaths for group A can be calculated as:

$$E_A = \sum E_{Ai} \tag{4}$$

The total number of the expected deaths in group B can be calculated based on the total number of expected deaths for group A given that the total number of events is n as follows:

$$E_B = n - E_A \tag{5}$$

In the Log-rank test, the data for the two groups combined are ordered and then each event, in turn, is considered starting at time $t = 0$. Then the log-rank statistics is calculated for two groups based on the summed observed minus expected score for a given group and its variance estimate and it is given as follows:

$$\chi^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \tag{6}$$

Here O_A and O_B represent the total number of events in groups A and B respectively and E_A and E_B represent the total number of expected events in group A and group B respectively. This statistic is compared with χ^2 statistics to decide whether there is a significant difference between the two groups or not using a specific confidence interval or level of significance.

2.5 Cox regression model

The Cox regression model also known as the Cox proportional hazards model (CPHM) is used to investigate the association between the survival time of patients and one or more predictor variables. CPHM is a regression model that has a dependent variable and independent variables and it is used to know the effect of specific variables on the event. Its formula is written as shown in the following equation:

$$h(t, X) = h_0(t)e^{(\sum_{i=1}^{p} B_i x_i)} \tag{7}$$

where $h_0(t)$ is the baseline hazard, $X's$ here are time-independent, and B_i are the regression coefficients. It is important to note that Kaplan-Meier curves and log-rank tests work with categorical predictor variable and they can describe the survival according to only one factor under investigation. CPHM model can work for quantitative predictor variables as well as categorical predictor variables and it can assess at the same time the effect of several factors on the survival time of patients.

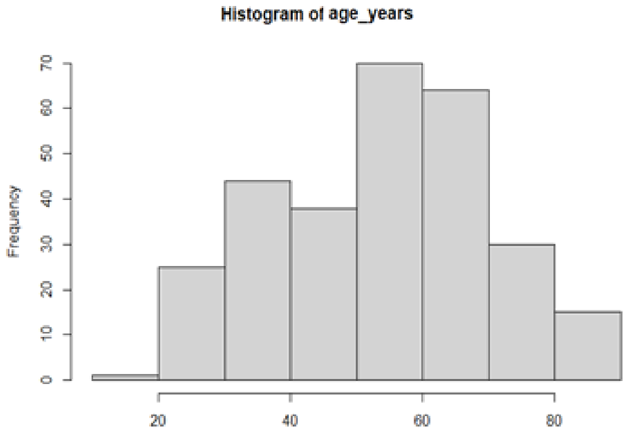

Figure 3: Histogram shows the distribution of age values.

3 Results and discussion

To analyze COVID-19 data, we used survival and survminer functions under R software. In this COVID-19 patients' data, the event of interest is the recovery of the patients and the outcome is time in days until the recovery. We must consider an important analytical problem called censoring, which occurs when we have sick people at risk who have died or the recovery time for them is not known due to losing their follow up, therefore the patients' exact recovery time will not be known at least in the period of the study so the patient survival time is considered censored. In other words, in this study, we will consider censoring if the patient is died or lost follow-up in the determined period given that the period of the study is from Jan 1 to Feb 11, 2020, as shown in the dataset subsection. After extracting the data and preparing it we read it in R software. We consider three predictive variables (patient gender, age, and time to hospitalization). The gender variable is a categorical variable and it can be easily analyzed using Kaplan–Meier survival curve and log-rank since we have two groups male and female. The patient age is a continuous variable, therefore, we need to convert it into categorical to be able to use it as a predictive variable. To do so we need to use a cutoff, where we will consider the age greater than this cutoff as old and the age less than the cutoff as young. To determine the cutoff, we should look at the overall distribution of age values using the histogram shown in Figure 3, where the cutoff of 50 is obviously suggested to be used. Also, we converted time to hospitalization to a categorical variable by considering hospitalization within 6 days as 'Early' and hospitalization in a time greater than 6 days as 'Late'.

Table 1: summary of the Kaplan-Meier estimates for the age groups up to day 16.

Old age group						
time	n. risk	n. event	non-recovery probability	std. err.	Lower. 95% CI	Upper. 95% CI
4	343	1	0.997	0.003	0.991	1.000
5	321	1	0.994	0.004	0.986	1.000
6	302	1	0.991	0.005	0.980	1.000
7	280	4	0.977	0.009	0.959	0.994
9	261	1	0.973	0.010	0.954	0.992
10	250	1	0.969	0.010	0.949	0.989
11	240	5	0.949	0.013	0.923	0.975
13	216	3	0.936	0.015	0.906	0.966
14	205	2	0.926	0.016	0.895	0.959
15	197	4	0.908	0.019	0.872	0.945
16	183	4	0.888	0.021	0.848	0.929
Young age group						
4	270	1	0.996	0.004	0.989	1.000
5	240	1	0.992	0.006	0.981	1.000
6	219	2	0.983	0.008	0.967	1.000
7	207	3	0.969	0.012	0.946	0.992
8	198	2	0.959	0.013	0.933	0.986
9	195	5	0.934	0.017	0.902	0.968
10	185	3	0.919	0.019	0.883	0.957
11	177	4	0.899	0.021	0.858	0.941
12	169	1	0.893	0.022	0.852	0.937
13	167	3	0.877	0.023	0.833	0.924
14	158	1	0.872	0.024	0.826	0.919
15	156	5	0.844	0.026	0.794	0.896
16	148	12	0.775	0.030	0.718	0.837

To analyze the data based on the age group, we created a survival object and we fit the Kaplan-Meier curves by passing the created survival object to survfit function. We obtained the results given in Table 1. Normally, the results obtained from the survfit function are the probability of non-recovery as shown in the 4th column of Table 1 i.e. death or negative event (Table 1 shows the results up to day 16). In this study, we are looking for a positive event (recovery) therefore we can calculate the recovery rate as (1- the probability of non-recovery). The results show that in the old age group over the four days period 1 recovered out of 343, therefore, the probability of non-recovery is (343-1)/343=0.997 (see Table 1 the first row) so the recovery rate is (1- 0.997) =0.003. Over the five days period as shown in the table (see Table 1 the second row), 21 patients of the remaining 342 patients lost follow-up (censored) so the number of remaining patients on the 5th day is 321. One of the remaining patients is recovered in the 5th day therefore, the proportion not recovered is 0.994. We could calculate the survival at a specific time t i.e $S(t) = p.1 * p.2 * \ldots * p.t$, where $p.1$ is the rate of the surviving patients who past the first time point and $p.2$ is the rate of the surviving patients who past the second time point, and so forth.

It is very important to take into account that starting from *p*.2 we should consider only those patients who survived past the previous time point to calculate the survival rate for the following time point, in other words, *p*.2, *p*.3..., *p*.*t* are survival rates that are conditional on the previous survival rates. Given the assumption of independent and random censoring, we assume that the 21 patients who were censored were similar to the 321 who remain at risk regarding their survival experience. Since 1 of the 321 who remained and survived on the 5th day recovered and we have 1 recovered on the 4th day then the total number of patients who recovered on the course of 5 days is 2. Subtracting 2 from the original number, which is 343 will yield 341. Then the recovery rate in the 5th day will be (1-341/343) = 0.006. The same analysis for the old age group is applied to the young age group (the results up to day 16 out of 30 days are shown in Table 1). The lower 95% confidence interval and upper 95% confidence interval tell us how accurate the estimate of the mean is [33]. In the first row in Table 1 the lower 95%CI and the upper 95%CI show us that we are 95% confident that the interval (0.991, 1.000) contains the true value of the parameter. Also, we can see that this interval is very narrow, which means that the certainty of the results is very high. In other words, we are 95% certain about the results. This narrow interval is associated with a very small standard error (0.003).

The corresponding survival curve can be obtained using the function ggsurvplot on the survival object. The obtained curves (see Figure 4) are step functions that allow us to compare the survival time of two age groups. Typically, the curve starts at 1 representing the fact that all of the patients are not having the event at entry into the study (see Figure 4 A). Over time the curve represents the probability of remaining non-recovered patients. Since we are interested in the probability of the recovered patients, we drew the survival curve starting from 0 to represent the portion of the recovered patients as shown in Figure 4 B. It is clear from Figure 4 B that the survival function of the young age group consistently lies above that for the old age group. This indicates that the young age group is better in recovering from COVID-19 than the old age group. We note that the two functions are somewhat close to each other in the first few days. This indicates that the young age group can survive COVID-19 later after infection than its early one. The estimate of the median recovery time for the young age group can be obtained from Figure 4 by selecting the value in the time axis that corresponds to the survival probability of 0.5. From the figure, it is clear that the median recovery time is greater than 20 days. The p-value of the log-rank is 0.00012, which indicates that the results are significant considering $p < 0.05$ indicates statistical significance, in other words the results show that there is a significant difference between young and old patients regarding the recovery from COVID-19.

To analyze the data based on the gender (Male, Female), we directly created a survival object since we don't need to convert the gender of the patient to a categorical variable (it is already a categorical variable). Then we fit the Kaplan-Meier curves by passing the created survival object to survfit function. We obtained the results given in Table 2 (we showed the results for the first 15 days). The results show that in the Female group over the four days period 1 recovered out of 295 therefore the probability of non-recovery is 294/295=0.997 (see the first row). Then the recovery rate in the 4th day will be (1-0.997) = 0.003. The rest of the Female group results and the male group results can be described as we did with the age group results that are given in Table 1.

Table 2: summary of the Kaplan-Meier estimates for the sex groups up to day 15.

			Female group			
time	n. risk	n. event	non-recovery probability	std. err.	Lower. 95% CI	Upper. 95% CI
4	295	1	0.997	0.003	0.990	1.000
5	261	3	0.985	0.007	0.971	1.000
6	234	1	0.981	0.008	0.964	0.998
7	216	4	0.963	0.012	0.939	0.987
8	210	1	0.958	0.013	0.933	0.984
9	202	5	0.934	0.016	0.903	0.967
10	193	3	0.920	0.018	0.885	0.956
11	184	5	0.895	0.021	0.855	0.937
13	169	2	0.884	0.022	0.842	0.928
15	160	1	0.879	0.022	0.836	0.924
			Male group			
4	346	2	0.994	0.004	0.986	1.000
5	325	1	0.991	0.005	0.981	1.000
6	310	3	0.982	0.007	0.967	0.996
7	293	4	0.968	0.010	0.949	0.988
8	283	1	0.965	0.010	0.944	0.985
9	275	2	0.958	0.012	0.935	0.981
10	262	1	0.954	0.012	0.931	0.978
11	252	4	0.939	0.014	0.912	0.967
12	240	1	0.935	0.014	0.907	0.964
13	233	4	0.919	0.016	0.888	0.952
14	220	4	0.902	0.018	0.868	0.938
15	210	8	0.868	0.021	0.828	0.910

The corresponding survival curve for the sex is shown in Figure 5, where the step functions allow us to compare the survival time of two sex groups. Figure 5 A is the probability of remaining unrecovered patients based on gender. The survival function of the Female group and that for the male group from the time 0 up to 40 follow similar paths, therefore the p-value (0.63) from the log-rank test is not significant considering $p < 0.05$ indicates statistical significance. Figure 5 B shows the survival curves starting from 0 and they represent the proportion of the recovered patients based on sex.

We used Cox regression model to measure the effect of the different factors on the recovery from COVID-9. in Cox regression the measure of the effect is hazard rate. The hazard is the instantaneous event rate or the probability of a patient at time *t* has an event at that time. Here the assumption is non-recovery if the event does not occur up to time *t* [23, 34]. Hazard ratio of 1 means that event rates are the same in the members of the same group. Figure 7 shows that at any time 1.9 as many patients in the young age group are having an event (recovery) proportionally to the old age group, which is taken as a reference, and the value 0.001** shows that this result is statistically significant. The result of the hazard ratios supports the results that we obtained in the step functions that are depicted in Figure 4 B. Regarding the sex group, the results in the figure shows that the hazard ratio is 1 which indicate that three is no difference between male patients and female patients in recovery from

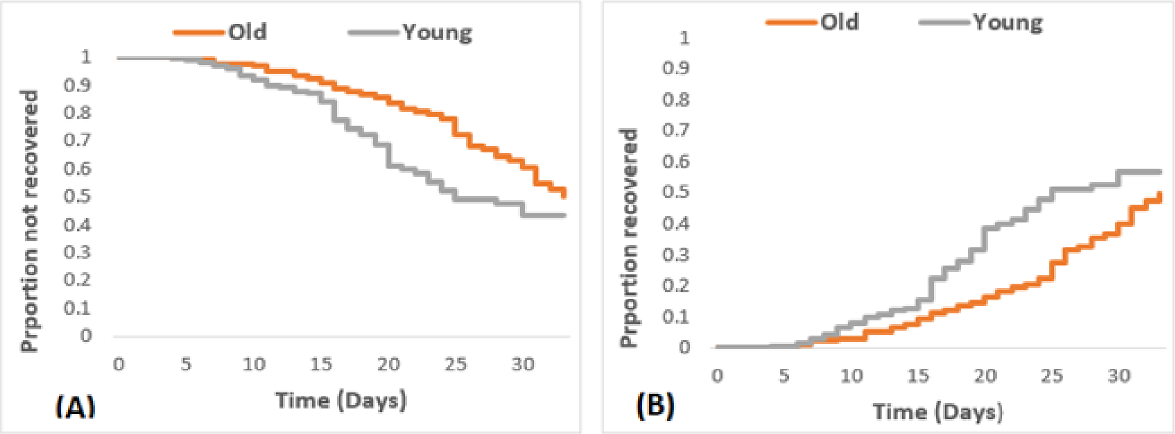

Figure 4: Survival curves for days to recovery from COVID-19 (age groups). A) Proportion not recovered; B) Proportion recovered.

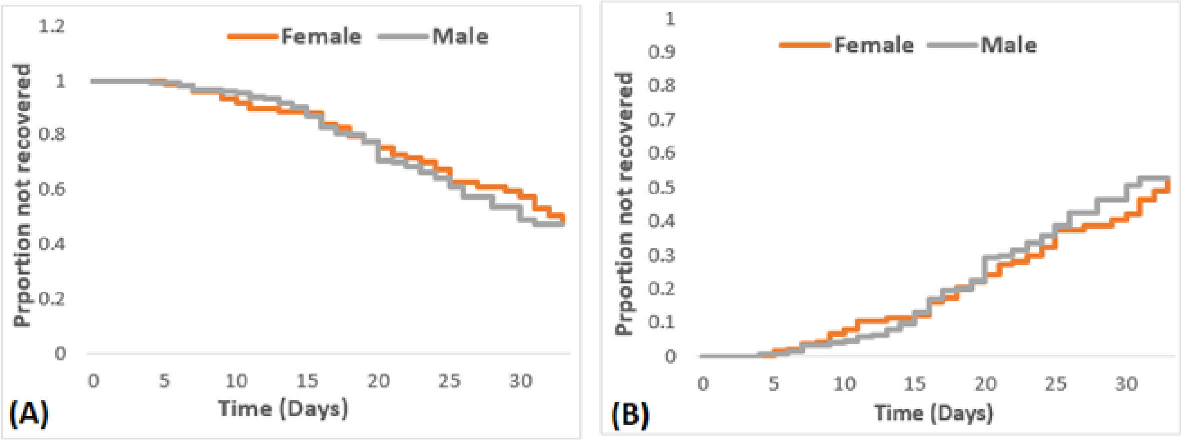

Figure 5: Survival curves for days to recovery from COVID-19 (sex groups). A) Proportion not recovered; B) Proportion recovered.

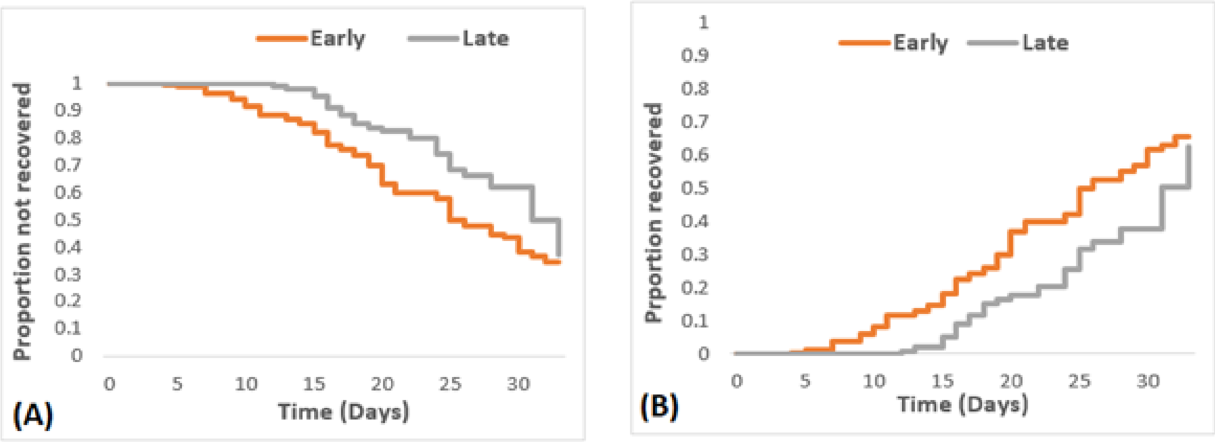

Figure 6: Survival curves for days to recovery from COVID-19 (time to hospitalization groups). A) Proportion not recovered; B) Proportion recovered.

COVID-19. This result supports the results that we obtained in the step functions. We note that the P-value is quite different from what is shown with the Kaplan-Meier estimator and the log-rank test that is because the hazard ratio calculates the hazard ratio and respective risk of death whereas Kaplan-Meier estimator and the log-rank test estimate the survival probability [35]. Therefore, we can see that the results yielded by these different methods are different in terms of significance.

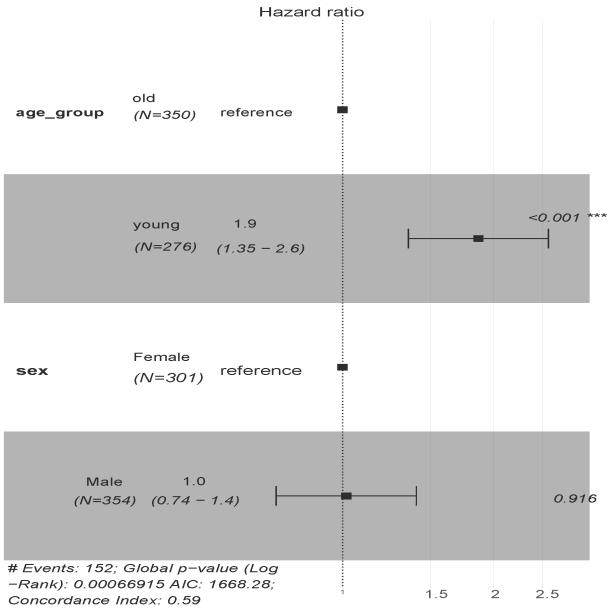

Figure 7: Forest plot shows the hazard ratios of the age and sex groups.

Also, we analyzed the data based on the time to hospitalization (early, late), we considered time to hospitalization as early if the patient is hospitalized within 6 days from catching the disease and as late if is hospitalized in a time greater than 6 days from catching the disease. Since not all the COVID-19 patients are hospitalized, we deleted the cases that have no hospitalization date. We then created a survival object and fit the Kaplan-Meier curves by passing the created survival object to survfit function. We obtained the results given in Table 3, which shows the results for the first 20 days. The results show that in the early time to hospitalization group over the four days period 1 recovered out of 159, therefore, the proportion not recovered rate is 158/159=0.994 (see Table 3 the first row), and therefore the proportion recovered rate is 1-0.994=0.006. In the late hospitalization group, the results show that over 12 days 1 recovered out of 103 so the proportion not recovered is 102/103=0.990 and hence the proportion recovered rate is 1-0.990=0.010. We note that in the early to hospitalization group the recovery starts at day 4, while in the late to hospitalization group the recovery starts at day 12.

The survival curve of the time to hospitalization groups is shown in Figure 6. It is clear from Figure 6 B that the survival function of the early time to hospitalization group consistently lies above that for the late time to hospitalization group. This indicates that the early time to hospitalization group is better recovering from COVID-19 than the late time to hospitalization group. Also, we note that the two functions are somewhat close to each other in the first few days (up to day 4). This indicates that the early time to hospitalization group can survive COVID-19 later after 4 days from infection than its early one. The p-value of the log-rank is 0.0052, which indicates that the results are significant considering $p < 0.05$ indicates statistical significance, in other words, the results show that there is a significant difference between the early time to hospitalization group and late time to hospitalization group.

Table 3: summary of the Kaplan-Meier estimates for the time to hospitalization groups.

Early time to hospitalization group						
time	n. risk	n. event	non-recovery probability	std. err.	Lower. 95% CI	Upper. 95% CI
4	159	1	0.994	0.006	0.981	1.000
5	157	1	0.987	0.009	0.970	1.000
7	147	4	0.961	0.016	0.930	0.992
9	133	3	0.939	0.020	0.901	0.978
10	124	3	0.916	0.023	0.872	0.963
11	118	4	0.885	0.027	0.833	0.940
13	110	2	0.869	0.029	0.814	0.928
14	103	2	0.852	0.031	0.794	0.915
15	100	4	0.818	0.034	0.754	0.887
16	92	5	0.774	0.037	0.704	0.851
17	85	2	0.755	0.039	0.683	0.835
18	81	2	0.737	0.040	0.662	0.819
19	77	4	0.698	0.042	0.620	0.786
20	70	7	0.629	0.046	0.545	0.724
Late time to hospitalization group						
12	103	1	0.990	0.010	0.010	0.972
13	100	1	0.980	0.020	0.014	0.954
15	96	3	0.950	0.050	0.022	0.908
16	89	4	0.907	0.093	0.030	0.851
17	82	2	0.885	0.115	0.033	0.823
18	74	3	0.849	0.151	0.037	0.779
19	68	1	0.837	0.163	0.039	0.764
20	63	1	0.823	0.177	0.040	0.748

Cox regression model for time to hospitalization yielded the hazard ratio, which represents relative that compares the early time to hospitalization group with the late time to hospitalization group as shown in Figure 8. A hazard ratio of 0.54 for the late hospitalization group tells us that patients who sent to hospital late have less opportunity of recovering compared to patients who sent to the hospital early, which served as a reference to calculate the hazard ratio. As shown by the forest plot, the respective 95% confidence interval is (0.35 – 0.84) and this result is significant (p-value=0.006). Using this model, we can see that the time to hospitalization variable significantly influences the patients' recovery from COVID-19. Also, We note that the obtained p-value is quite different from what is shown with the Kaplan-Meier estimator and the log-rank test and that is due to the same justification that we presented when analyzing the sex and age groups.

Salinas-Escudero et al. study [24], which applied Kaplan-Meier and Cox regression models to the Mexican found that the age factor has a significant effect in recovering from COVID-19. This finding agrees with our finding on the data we used. In another hand, their study found that sex group has significant effects, which disagrees with our finding. Monira Mollazehi et al study [25] applied Weibull model in Singapore. The factors they used are age and nationality. Their finding agrees with ours regarding the age group.

The limitations that need to be declared in this research are: First, the dataset is for a specific region and in a specific period. Second, the dataset is relatively small compared to the total infected cases.

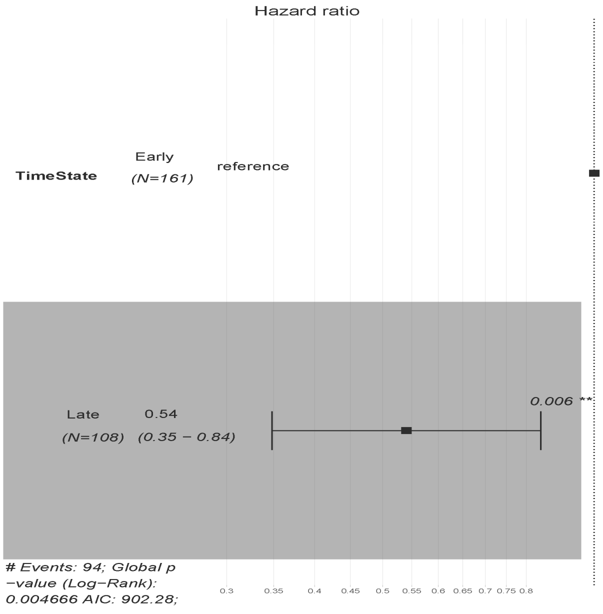

Figure 8: Forest plot shows the hazard ratios of time to hospitalization group.

4 Conclusion

In this work, we used survival analysis to analyze COVID-19 data that we obtained from the clinically diagnosed and confirmed cases where the date of onset is recorded in Wuhan and elsewhere in China from Jan 1 to Feb 11, 2020. We used the Kaplan-Meier method which is an estimator of survival probabilities and the Cox regression model, which is known as the Proportional Hazard Model for the analysis. The event of interest in our analysis is the recovery of the patients from COVID-19 and the outcome is time in days until the recovery. The predictor variables that we used are sex, age, and time to hospitalization. The results show that the young age group is better in recovering from COVID-19 than the old age group with a significant difference (P-value = 0.00012) and at any time 1.9 as many patients in the young age group is having an event (recovery) proportionally to the old age group. The step functions of the sex group show that the female and male groups are somewhat close to each other in recovering from COVID-19 and the p-value =0.63 indicates that there is a non-significant difference in the results between Male and Female considering $p < 0.05$ indicates statistical significance. The results also show that early time to hospitalization group can recover from COVID-19 better than late time to hospitalization group (the p-value of the log-rank is 0.0052)

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors would like to acknowledge the Jouf University for all the support that it provides.

References

[1] M. Cascella, M. Rajnik, A. Cuomo, S. Dulebohn, N. R. Di, "Features, evaluation and treatment coronavirus (COVID-19)," Statpearls [internet]: StatPearls Publishing, 2020.

[2] N. Zhu, D. Zhang, W. Wang, X. Li, Y. et al., "A novel coronavirus from patients with pneumonia in China, 2019," New England Journal of Medicine, 2020, doi:10.1056/NEJMoa2001017.

[3] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. et al., "A new coronavirus associated with human respiratory disease in China," Nature, **579**(7798), 265–269, 2020, doi:10.1038/s41586-020-2008-3.

[4] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, Z. et al., "A pneumonia outbreak associated with a new coronavirus of probable bat origin," nature, **579**(7798), 270–273, 2020, doi:10.1038/s41586-020-2012-7.

[5] V. M. Corman, D. Muth, D. Niemeyer, C. Drosten, "Hosts and sources of endemic human coronaviruses," Advances in virus research, **100**, 163–188, 2018, doi:10.1016/bs.aivir.2018.01.001.

[6] A. E. Gorbalenya, L. Enjuanes, J. Ziebuhr, E. J. Snijder, "Nidovirales: evolving the largest RNA virus genome," Virus research, **117**(1), 17–37, 2006, doi:10.1016/j.virusres.2006.01.017.

[7] M. K. Elbashir, M. Ezz, M. Mohammed, S. S. Saloum, "Lightweight Convolutional Neural Network for Breast Cancer Classification Using RNA-Seq Gene Expression Data," IEEE Access, **7**, 185338–185348, 2019, doi: 10.1109/ACCESS.2019.2960722.

[8] A. Ancarani, C. Di Mauro, L. Fratocchi, G. Orzes, M. Sartor, "Prior to reshoring: A duration analysis of foreign manufacturing ventures," International Journal of Production Economics, **169**, 141–155, 2015, doi:10.1016/j.ijpe.2015.07.031.

[9] E.-Y. Jung, C. Baek, J.-D. Lee, "Product survival analysis for the App Store," Marketing Letters, **23**(4), 929–941, 2012, doi:10.1007/s11002-012-9207-0.

[10] N. E. Buckley, P. Haddock, R. D. M. Simoes, P. et al., "A BRCA1 deficient, NFκB driven immune signal predicts good outcome in triple negative breast cancer," Oncotarget, **7**(15), 19884, 2016, doi:10.18632/oncotarget.7865.

[11] S. R. Gross, B. O'Brien, C. Hu, E. H. Kennedy, "Rate of false conviction of criminal defendants who are sentenced to death," Proceedings of the National Academy of Sciences, **111**(20), 7230–7235, 2014, doi:10.1073/pnas.1306417111.

[12] M. A. Alvi, D. G. McArt, P. Kelly, M.-A. Fuchs, A. et al., "Comprehensive molecular pathology analysis of small bowel adenocarcinoma reveals novel targets with clinical utility," 2015, doi:10.18632/oncotarget.4576.

[13] D. Murray, A. Carr, C. Bulstrode, "Survival analysis of joint replacements," The Journal of bone and joint surgery. British volume, **75**(5), 697–704, 1993, doi:10.1302/0301-620X.75B5.8376423.

[14] T. Sørlie, C. M. Perou, R. Tibshirani, A. et al., "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," Proceedings of the National Academy of Sciences, **98**(19), 10869–10874, 2001, doi:10.1073/pnas.191367098.

[15] F. Emmert-Streib, M. Dehmer, Medical biostatistics for complex diseases, John Wiley & Sons, 2010.

[16] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu, R. Kuang, "Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment," PLOS Computational Biology, **9**(3), e1002975, 2013, doi:10.1371/journal.pcbi.1002975.

[17] M. J. Campbell, T. D. V. Swinscow, Statistics at square one, John Wiley & Sons, 2011.

[18] M. J. Hancock, C. G. Maher, L. d. C. M. Costa, C. M. Williams, "A guide to survival analysis for manual therapy clinicians and researchers," Manual therapy, **19**(6), 511–516, 2014, doi:10.1016/j.math.2013.08.007.

[19] S. J. Staffa, D. Zurakowski, "Competing risks analysis of time-to-event data for cardiovascular surgeons," The Journal of thoracic and cardiovascular surgery, **159**(6), 2459–2466, 2020, doi:10.1016/j.jtcvs.2019.10.153.

[20] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. Voelker, B. Nussenbaum, E. W. Wang, "A practical guide to understanding Kaplan-Meier curves," Otolaryngology—Head and Neck Surgery, **143**(3), 331–336, 2010, doi: 10.1016/j.otohns.2010.05.007.

[21] G. David, K. Kleinbaum, Survival analysis: a self-learning text, Springer-Verlag New York, 2016.

[22] E. L. Kaplan, P. Meier, "Nonparametric estimation from incomplete observations," Journal of the American statistical association, **53**(282), 457–481, 1958, doi:10.2307/2281868.

[23] D. R. Cox, "Regression models and life-tables," Journal of the Royal Statistical Society: Series B (Methodological), **34**(2), 187–202, 1972.

[24] G. Salinas-Escudero, M. F. Carrillo-Vega, V. Granados-García, S. Martínez-Valverde, F. Toledano-Toledano, J. Garduño-Espinosa, "A survival analysis of COVID-19 in the Mexican population," BMC public health, **20**(1), 1–8, 2020, doi:10.1186/s12889-020-09721-2.

[25] M. Mollazehi, M. Mollazehi, A.-S. G. Abdel-Salam, "Modeling Survival Time to Recovery from COVID-19: A Case Study on Singapore," 2020, doi: 10.21203/rs.3.rs-18600/v1.

[26] Q. Wang, S. Xie, Y. Wang, D. Zeng, "Survival-Convolution Models for Predicting COVID-19 Cases and Assessing Effects of Mitigation Strategies," medRxiv, 2020, doi:10.1101/2020.04.16.20067306.

[27] N. Barda, D. Riesel, A. Akriv, J. Levy, U. Finkel, G. Yona, D. Greenfeld, S. Sheiba, J. Somer, E. Bachmat, et al., "Developing a COVID-19 mortality risk prediction model when individual-level data are not available," Nature communications, **11**(1), 1–9, 2020, doi:10.1038/s41467-020-18297-9.

[28] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, W. et al., "Estimates of the severity of coronavirus disease 2019: a model-based analysis," The Lancet infectious diseases, 2020, doi:10.1016/S1473-3099(20)30243-7.

[29] M. Pampaka, G. Hutcheson, J. Williams, "Handling missing data: analysis of a challenging data set using multiple imputation," International Journal of Research & Method in Education, **39**(1), 19–37, 2016, doi:10.1080/1743727X.2014.979146.

[30] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, et al., "The prevention and treatment of missing data in clinical trials," New England Journal of Medicine, **367**(14), 1355–1360, 2012, doi:10.1056/NEJMsr1203730.

[31] J. W. Graham, P. E. Cumsille, A. E. Shevock, "Methods for handling missing data," Handbook of Psychology, Second Edition, **2**, 2012, doi: 10.1002/9781118133880.hop202004.

[32] D. G. Kleinbaum, M. Klein, Survival analysis, Springer, 2010.

[33] J. H. McDonald, Handbook of biological statistics, volume 2, sparky house publishing Baltimore, MD, 2009.

[34] D. W. Hosmer Jr, S. Lemeshow, S. May, Applied survival analysis: regression modeling of time-to-event data, volume 618, John Wiley & Sons, 2011.

[35] N. H. Ng'andu, "An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model," Statistics in medicine, **16**(6), 611–626, 1997, doi:10.1002/(sici)1097-0258(19970330)16:6⟨611::aid-sim437⟩3.0.co;2-t.