Survival analysis on the mortality of covid

Do Thanh Dat LE, Piseth KHENG

December 9, 2023

1 Introduction

During the latter of 2019, the world was hit by a sudden and widespread outbreak of COVID-19, a disease that originated in Wuhan, China. It quickly spread across the globe, posing one of the biggest challenges humanity has faced in recent times. COVID-19 is highly contagious and can spread through the air from person to person, making it difficult to prevent contact between infected and non-infected individuals. Due to the challenges posed by this disease, the death rate among infected people has been increasing steadily.

In our project, We used the Cox proportional hazards model to analyze the relationship between patient variables, such as age, symptoms, and chronic diseases, and time to death. The dataset we received was not initially structured for survival analysis. Therefore, our primary task was to clean the data, which is a critical step in ensuring precise analysis. We extracted relevant features and eliminated irrelevant ones from the original dataset for our study. We then used graphical and numerical summaries to explore the features of the dataset and check the survival function in the subgroups defined by each categorical variable.

Next, we applied two types of the Cox model to our dataset: one with a linear relationship and the other with a non-linear relationship. In addition, we also intended to use the Random Forest machine learning model, which is a tree-based model, and compare its performance with the Cox model to select the better one for our project.

2 Exploratory data analysis

2.1 General information

The dataset is a subset of a sample extracted from administrative data on confirmed COVID cases collected worldwide in the early stage of the pandemic. The essential variables in the dataset include: `date_onset_symptoms`, `date_admission_hospital`, `symptoms`, `outcome` (death or discharge), `date_death_or_discharge`, `age`, `sex`, `chronic_disease_binary`, `chronic_disease`, `latitude`, `longitude`, and `country`.

As previously mentioned, the dataset was not initially designed for survival analysis and lacks important variables such as the `Duration` (time to death caused by disease) and `censoring indicator`. Additionally, the dataset contains a variety of data types including numerical, categorical, string, and date formats.

2.2 Data preprocessing

In this section, we perform several important data preprocessing tasks. These include encoding duration and censoring indicator, encoding hospitalization information, extracting relevant information from the symptoms feature, and correcting any typos in the data.

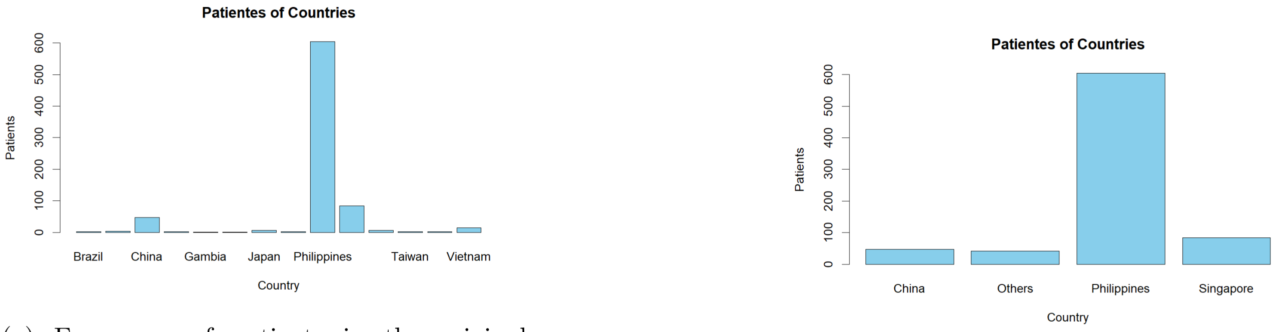
We begin by encoding the variables of interest, namely the `duration` and `censoring indicator`. The duration variable represents the time elapsed between the onset of symptoms and the date of death or discharge. However, we noticed that some duration values are negative, which we traced to errors in the recorded data between the `"date_onset_symptoms"` and `"date_death_or_discharge"` fields. To address this issue, we assume that the recording date in "date_onset_symptoms" was incorrect and should have been in the "date_death_or_discharge" field, and vice versa. Consequently, we convert the negative values in the "Duration" variable to positive values.

Next, we need to encode the censoring indicator from the outcome variable. In this case, the censoring indicator variable will be set to 1 when the patient experiences the ”death” event and 0 when the patient has the ”discharge” event. However, we have come across some issues with the outcome data as it is not uniquely defined. To address this, we have converted the data into the unique values of ”death” and ”discharge” before we proceed with the encoding process. We aim to encode hospitalization information from the date of admission variable and add a new variable that represents the time elapsed between the onset of symptoms and the date of hospitalization admission, namely time_until_hospitalization.

Extracting information from the variable symptoms was the most difficult part of the process. This was due to the fact that there was less recorded data on symptoms compared to the total number of patients. The first step was to identify the number of unique symptoms present in the dataset, which was found to be around 50, including some typos. After correcting these errors, we had to group similar symptoms together to reduce the number of features as some symptoms were recorded for only one patient. We have grouped together 12 symptoms based on the Centers for Disease Control and Prevention (CDC) [1]. These symptoms include fatigue and weakness, difficulty breathing, chest discomfort, muscle aches, cough, throat symptoms, fever and chills, neurological symptoms, gastrointestinal symptoms, respiratory symptoms, pneumonia, and severe symptoms.

After we have grouped similar symptoms, we will convert them into binary variables. These variables will have a value of 1 if the patient has that symptom and 0 if the patient does not have it. In total, there will be 12 binary variables corresponding to the 12 symptoms being considered.

Last but not least, we also take into consideration some other important factors such as age, chronic diseases, and country. Chronic diseases can exacerbate the condition of COVID-19 patients. The chronic disease variable is represented by a string that contains the name of the disease. Based on this, we create a variable called ”number_chronic_disease”, which represents the number of chronic diseases that each patient has. However, we have also come across some unusual values in the age variable, where the age interval (i.e., 50-59) has been recorded. Therefore, we attempt to transform those values to the middle age of that interval (i.e. 55).



(a) Frequency of patients in the original dataset

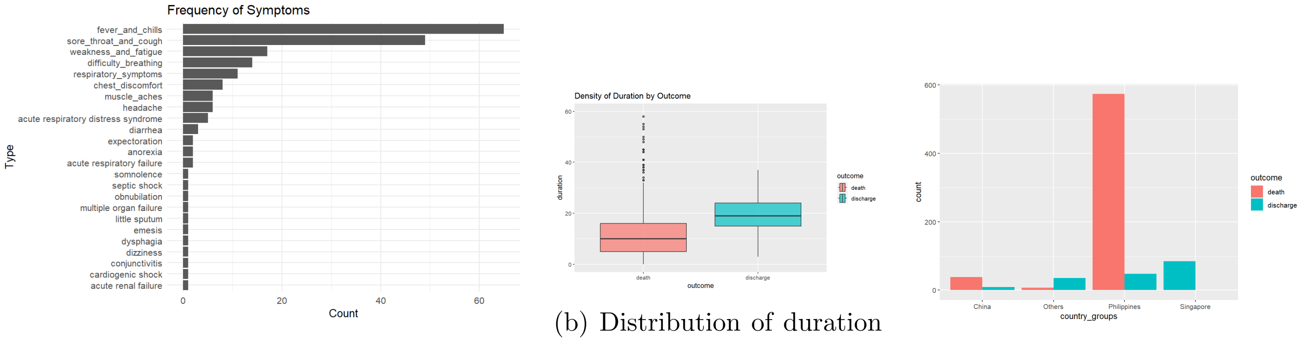
(b) Frequency of patients after transformed

The data shown in figure 1a indicates that the Philippines has the highest number of patients, while many other countries have lower frequencies that may not be significant for our analysis. Thus, we propose grouping these countries together, as shown in figure 1b.

2.3 Data analysis

After finishing the data cleaning procedure, we will analyze categorical features to increase precision when conducting survival analysis.

According to the bar chart in figure 2a, it appears that the most frequent symptoms are fever_and_chills and sore_throat_and_cough. Additionally, some symptoms such as weakness and fatigue and difficulty



(a) Frequency of each symptoms

(b) Distribution of duration by outcome

(c) Frequency of patients

`breathing` are moderately frequent. Moreover, the frequency of some symptoms is extremely low, and we therefore believe that they have no impact on our analysis. Moreover, as shown in figure 2b, We observed that patients who died had a shorter duration of hospitalization compared to survivors; however, there are some cases where death occurred after a longer duration. Regarding the figure 2c, The graph shows that the Philippines has the highest number of patients compared to other countries, and it also appears to have the highest death rate. China also has a high death rate, but the total number of patients in China is only around 50. Surprisingly, there were no reported deaths in Singapore during the study. For the other countries, the death rate seems to be extremely low.

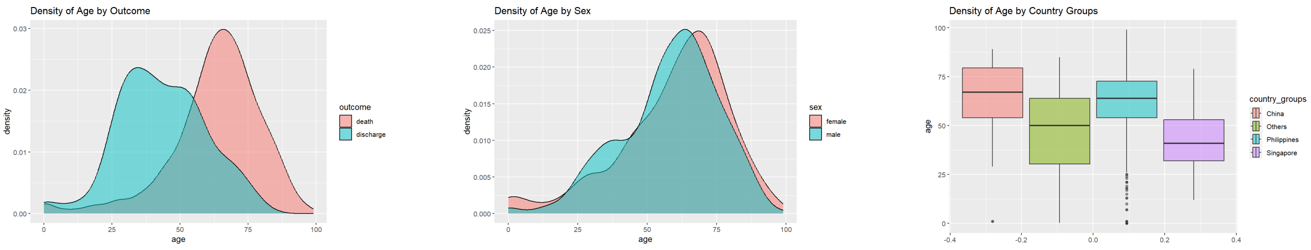


Figure 3: Distribution of patient age

Based on the graph 3, it appears that older people have a higher death rate compared to younger individuals. Additionally, when we consider the 'sex' feature, both male and female patients seem to have similar population density. Furthermore, a significant number of patients aged over 50 are from China and the Philippines.

3 Statistical modeling and survival analysis

After data preprocessing and data analysis, we continue with the modeling and survival analysis.

3.1 Data preparation

We split the data into train data (75%) and test data (25%). The train data was used to train the models and the test data was used to evaluate the models. The split was stratified well on the censorship, the percentage of censors in the train data is 77.36% and the percentage of censors in the test data is 77.32%.

We noticed that the covariate `hospitalized` is the time-dependent covariate. Therefore, we transformed the train data into a start-stop format to fit with the start-stop Cox model for this time-dependent covariate.

3.2 Start-stop Cox model

After fitting with the start-stop Cox model for all covariates. In the start-stop COX model, only `travel_history_binary` is statistically significant in the model with a significant level $\alpha = 0.05$. The coefficients of three covariates: `throat_symptoms`, `hospitalized`, and `country_groups` may be infinite with very high p-value, so these covariates are not statistically significant in the start-stop model.

Therefore, we will eliminate these 3 covariates out of the start-stop Cox model. Since the time-dependent covariate hospitalized was eliminated, we tried the start-stop model and the old format Cox model (without 4 covariates hospitalized, time_until_hospitalization, throat_symptoms, and country_groups). Then we used stepwise variable selection with criteria AIC to choose the best model with the minimum AIC.

3.3 Cox model with linear relationship

After the variable selection with the stepwise method with criteria AIC, we observed that the old format model was the same model as the start-stop model. This is because the time-dependent covariate hospitalized was removed from the model.

Covariates	exp(coef)	p-value
age	1.008	0.008
travel_history_binary	0.072	< 0.001
latitude	1.026	0.009
longitude	1.056	< 0.001
difficulty_breathing	0.368	0.021
cough_symptoms	0.602	0.133
gastrointestinal_symptoms	0.161	0.014
pneumonia_symptoms	14.671	< 0.001

Table 1: Cox model with linear relationship covariates

We obtained a model with 8 covariates: age, travel_history_binary, latitude, longitude, difficulty breathing, cough_symptoms, gastrointestinal_symptoms, pneumonia_symptoms. Moreover, with a significant level $\alpha = 0.05$, the cough_symptoms was not statistically significant, other covariates were statistically significant. The hazard ratio of age is greater than 1, indicating that when the age increases by 1 year, the risk of death increases 0.8%. Hence, old COVID-19 patients have a higher death risk than young patients. The hazard ratio of travel_history_binary is significantly less than 1 (0.072), so if the patients have travel history, the risk of death decreases significantly. The hazard ratios of 2 symptoms ”difficulty in breathing” and ”gastrointestinal symptoms” are less than 1, so the patients with these symptoms have lower death risk than the patients without. The hazard ratio of the pneumonia symptoms is much greater than 1, which means the pneumonia symptoms will highly increase the risk of death in COVID-19 patients. Then we check for linearity with martingale residuals

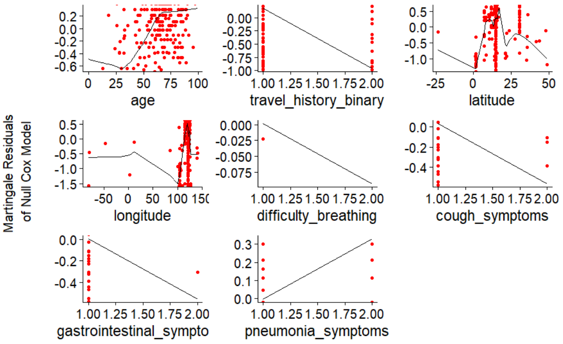
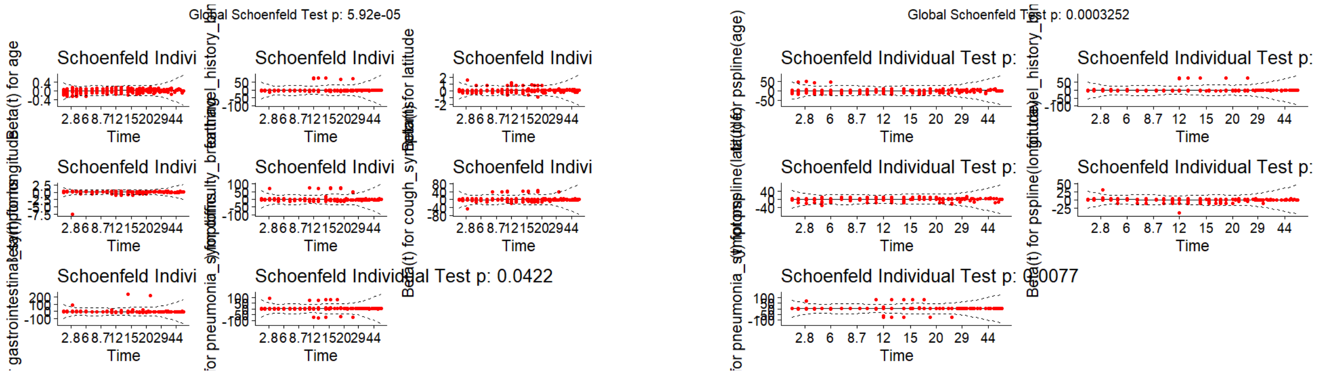


Figure 4: Martingale residuals for each covariate

We only focus on the three continuous covariates (age, latitude, longitude) in the Figure 4. These covariates have nonlinear martingale residuals. Then we try the Cox model including the nonlinear relationship of these covariates.

3.4 Cox model with nonlinear relationship

After fitting the Cox model with a nonlinear relationship to the train data, we used stepwise variable selection with criteria AIC to obtain the best model with minimum AIC. The model includes the 2 components (linear and nonlinear) of 3 covariates age, latitude, longitude, and 2 binary covariates travel_history_binary and pneumonia_symptoms. For significant level $\alpha = 0.05$, the covariates travel_history_binary, pneumonia_symptoms and the nonlinear component of longitude is significant. We check the proportional hazards assumption (time invariance) via Schoenfeld residuals for linear and nonlinear model



(a) Schoenfeld residuals for linear Cox model

(b) Schoenfeld residuals for nonlinear Cox model

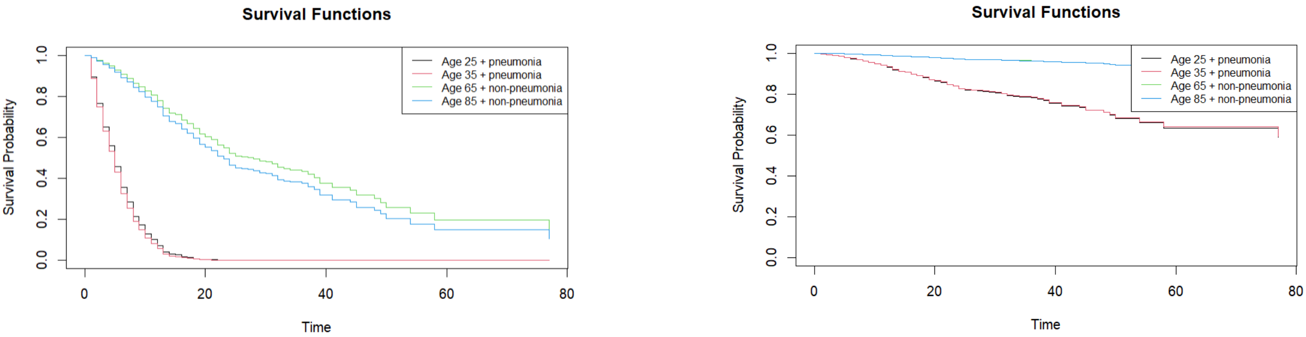
Figure 5: Schoenfeld residuals test

For the Cox model with linear relationship, the global Schoenfeld test has p-value ¡ 0.05, indicating that at least one covariate violates the assumption (its coefficient is not constant over time). The covariates longitude and gastrointestinal_symptoms have p-value > 0.05 so they have constant coefficients. Other covariates with p-value ¡ 0.05, indicating that their coefficients are not constant over time. For the Cox model with nonlinear relationship, only covariate longitude has a constant coefficient (p-value > 0.05).

3.5 Ages and pneumonia symptoms

In those 2 models above, we can see that the age and pneumonia symptoms have effects on COVID-19 patients, then we will analyze these 2 factors.

We defined a set of 4 patients with different ages, young patients (age 25, 35) have pneumonia symptoms, and old patients (age 65, 85) do not have pneumonia symptoms. All patients do not have other symptoms and travel history. Then we estimate the survival functions for 4 patients using 2 Cox models above. The results are represented in the graphs below



(a) Cox model with linear relationship

(b) Cox model with nonlinear relationship

Figure 6: Survival functions estimation by Cox models

Regarding the graph above, the pneumonia symptoms have a strong negative effect on the survival probability of COVID-19 patients over time. The survival probability of young patients (age 25, 35) with pneumonia symptoms decreases even faster than the survival probability of old patients (age 65, 85) without pneumonia symptoms. Therefore, the pneumonia symptoms have a much higher negative effect on the COVID-19 patients compared to the age.

Furthermore, we assume that the first patient (age 25 + pneumonia symptoms) from the patients above has been alive for 7 days, then we estimate the probability that this patient will be still alive for another 14 days (still alive on the day 21st) using the Cox model with a linear relationship and Cox model with nonlinear relationship. We obtain the results in the table below

Patients	Cox model linear	Cox model nonlinear
age25 + pneumonia	0.37%	85.5%
age25 + pneumonia + alive 7 days	1.3%	88.7%

Table 2: Probability of the patient still alive on the day 21st

From the table above, we could draw a conclusion that if a patient has been alive for 7 days, he/she will have a higher chance of being still alive for another 14 days.

3.6 Tree-based model using random forest

We fit the Random Forest model by rfsrc function in library randomForestSRC with default parameters, then we use grid search to optimize the parameters: ntree (number of trees), mtry (number of variables to possibly split at each node), nodedepth (maximum depth of a tree) to got the optimal model.

3.7 Model comparison using Brier score

In order to compare between models, we used the 3 models: Cox model linear, Cox model nonlinear, and Random Forest to predict the test data, then we computed the Brier Score. The results are represented in the table below

Model	Brier Score
Cox model with linear relationship	0.194
Cox model with nonlinear relationship	0.181
Random Forest with optimal parameters	0.179

Table 3: Brier score of three models

The model with a lower Brier score is better. Regarding the results in the table above, the Brier score of the Random Forest model with optimal parameters is the lowest so Random Forest model is best model among 3 models in predicting the survival probability of COVID-19 patients

4 Conclusions

In conclusion, the Random Forest model with optimal parameters is the most effective model for predicting the survival probability of COVID-19 patients based on the Brier score. However, this model is more complex to interpret compared to the two Cox models. All three models have Brier scores below 0.25, indicating their usefulness in prediction. If the goal is to better understand the impact of covariates on the survival probability of COVID-19 patients, the two Cox models would be a better choice.

References

[1] CDC. Symptoms of covid-19. 2022.