Critcare data analysis

Robert Farry - 200450971

May 5th 2023

Keywords— Co-variate(s), significant, analysis

Abstract

In this paper we're going to be looking at a sample of data taken from a subset of a larger set of data on survival of COVID-19 patients receiving critical care and we'll analyse this data in many different ways using non-parametric and semi-parametric methods with the aim of coming to some conclusions about which model best explains the data.

Contents

1 Overview of the Critcare data

Before getting into the main analyses we would like to become familiar with our data. Over this section I'm going to consider many different aspects of the data and look at the explanatory variables with a goal in mind of identifying different characteristics of each variable.

One of the main things we notice upon inspection of our data is the use of right censoring in the sample. This negates any need to concern ourselves about a patients start time (all start at the same time), instead we can focus on time of event (before the study finishes) and on censoring.

Another key feature of this sample is the size of it. We have 600 patients in our sample which is a good size. A large sample size gives us room to make significant assumptions in our analysis.

1.1 Defining our variables in this overview

In our data we have a combination of discrete and continuous explanatory variable. In the interest of convenience we shall define each variable under whichever group it suits best according to it's characteristics, i.e. a variable like BMI has an appearance of continuity but in reality is actually discrete, therefore to aid our analysis we shall treat it as continuous. It's very difficult to group discrete data that increased by small increments

1.2 The "continuous" variables

In this section we will cover variables that we have defined as "continuous" for the sake of convenience in this analyses even though they may not be strictly continuous by common standards.

Variable	**Mean**	**Range**	**Standard deviation**
Age	58	(18,97)	15.2
BMI	31.0	(15.1,44.8)	6.8
Apache II Score	29	(5,57)	11.1

1.2.1 Age

We find the majority of people in critical care are between ages 40 and 80 ($Pr(40 < Age < 80) = 0.88$) This makes sense in the context of the study because COVID-19 tends to affect people in the later stages of life more than those who are younger.

1.2.2 BMI

We see in the data though that 80% of the sample have a BMI over 25 and roughly half of the sample have a BMI over 30. If BMI > 25 this often indicates a person is overweight and BMI $> 30 \implies$ obesity. It's probably safe to say then that most of our sample struggle with issues of being overweight/obese (unless they have lots of muscle mass, which is unlikely amongst older people)

1.2.3 Apache II Score

According to medscape.com [1], ”The accuracy of the APACHE II at admission as an early prognostic indicator of disease severity is about 75%”, we can theorise then that this variable should be of great importance in our analysis. especially given the high mean for a categorical variable with the range (0,71).

1.3 The discrete variables

Now we've dealt with the ”continuous” variables we shall deal with the discrete variables and give an overview for each.

Variable	**binary?**	**1 count**	**proportion compared to full sample**
Gender (female = 1)	yes	197	32.8%
Comorbidities	yes	57	9.5%
Invasive Ventilation	yes	329	54.8%
Dependency	yes	50	8.3%
Deprivation	no	N/A	N/A
Status	yes	257	42.8%

1.3.1 Status

In this critical care data the type of censoring employed is right censoring. In our data 1 = death, 0 = censored. roughly half of our data not censored, this is a fairly standard rate for survival data.

1.3.2 Gender

A third of our data are female, we would probably expect an even split in the genders. A split like this then may be significant in our analysis.

1.3.3 Comorbidities

only about a tenth of our data have comorbidities present so this may not be very significant in our analysis.

1.3.4 Invasive Ventilation

About half of our data needed invasive ventilation within the first 24 hours of critical care so this may be quite significant in our analysis.

1.3.5 Dependency

Dependency is quite low here with one twelfth of the data falling under the category of dependent. This variable may not be significant in our analysis then.

1.3.6 Deprivation

Category	Count	Proportion
1	81	13.5%
2	97	16.2%
3	122	20.3%
4	150	25%
5	150	25%
Total	600	100%

We can see here that half of our sample are in the highest categories for the deprivation index. We can infer then that this variable will probably be quite significant in our analysis later on.

1.4 Survival function, S(t)

1.4.1 30-Day Survival Probability

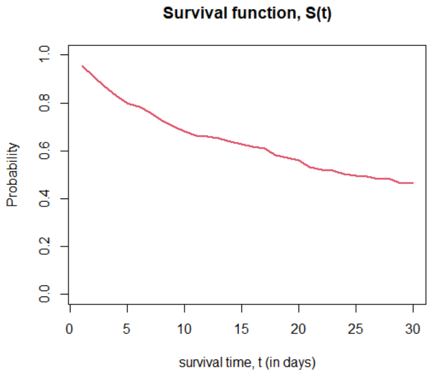

Figure 1: Survival function for our data

We can see here that the our survival function only decreases half way after 30 days (S(30) = 0.464). We also know that the median survival time is roughly 24 days. This could indicate that this sample needs a larger event time, T.

1.4.2 Time

Our times are ∈ (0,30). 30 days, on the surface, seems like a good length of time for critical care data as you would expect someone to die within this time if they're incredibly ill. However, we see 69 censors at t = 30 (end of experiment), this may mean that the sample taken isn't the most appropriate for this type of data as t should be a higher value so that we lose less information.

2 The survival experience of patients depending on selected risk factors

2.1 Introduction

In this section we're going to be taking various risk factors from the data and comparing patients with and without these risk factors, hoping to see some differences in order to decide whether these variables are significant in our analysis.

2.2 Survival experiences of patients:

2.2.1 With and without dependencies

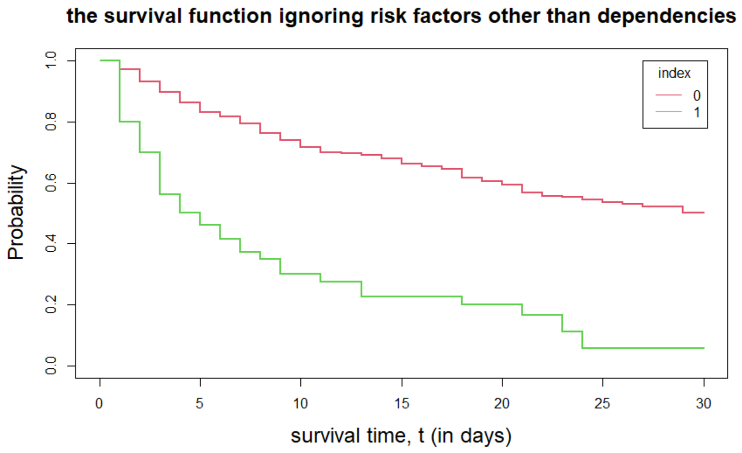

Figure 2: Survival function for patients with and without dependencies

From looking at figure 2, we can visually see that there is quite a significant difference in survival times for patients with and without dependencies. a patient with dependencies has a median survival time of 4 days, whereas a patient without dependencies has a median survival time of 29 days, these are some very significant differences.

If we compute a log-rank test in R, comparing the 2 survival curves for dependencies, we obtain a p-value of 2e-16. This is abundantly smaller than 0.05 so we can say there is extremely strong evidence that prognosis is worse in groups with dependencies.

2.2.2 With and without invasive ventilation

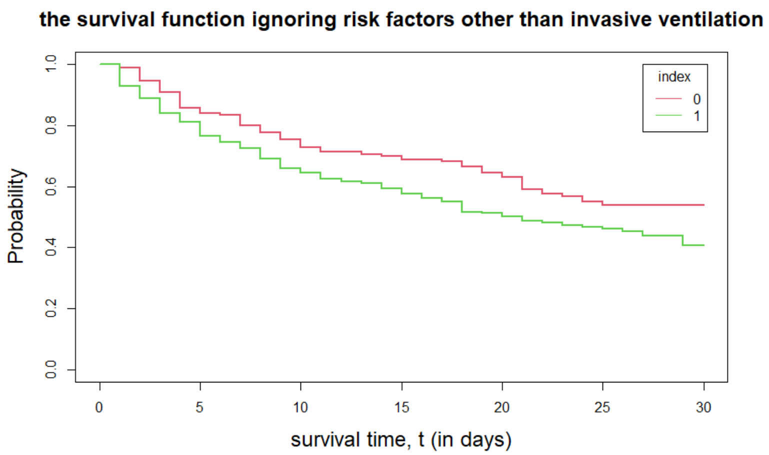

Figure 3: Survival function for patients with and without invasive ventilation

In figure 3 we can see that there is definitely a difference between patients with and without invasive ventilation. The difference looks relatively small as the curves are quite close together. However, the curves never intercept so it can be inferred that there is still a clear difference between these risk factors.

Once again, computing a log-rank test in R allows us to make some more accurate distinctions between the two curves and decide whether this risk factor is significant in our analysis. Doing this test in R, we obtain a p-value of 0.004 which is considerably smaller than 0.05. Therefore, there is strong evidence to suggest that prognosis is worse in patients with invasive ventilation compared to those without it.

2.2.3 With and without Comorbidities

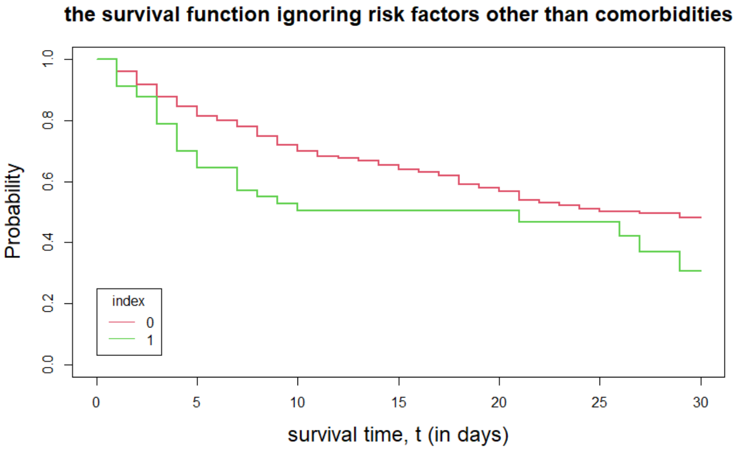

Figure 4: Survival function for patients with and without Comorbidities

In this figure we can see a difference between patients' survival times with and without comorbidities. The curves are quite close together so it may be a relatively small difference though.

Computing a log-rank test in R we obtain a p-value for this co variate of 0.01 (<0.05). Therefore we have evidence to suggest this co variate is significant but perhaps not as significant as dependencies or invasive ventilation.

3 Cox proportional hazards fit

To further our analyses of the critcare data, we are going to consider a cox proportional hazards fit. We'll do this in order to test which factors contribute most to our analysis in order to form a model that'll hopefully grant us some insights into the data. The covariates we'll be using for our fit are Age, Gender and BMI.

3.1 Analysis of our covariates

Factor	**exp(coefficient)**	**P-value**
Age	1.056509	2e-16
Female	0.835422	0.188
BMI	1.050297	1.98e-07

The table above tells us which of our covariates are significant. Both age and BMI have p-values significantly smaller than 0.05, whereas female has quite a large p-value. We can therefore deduce that female doesn't have strong evidence therefore it's not significant in our analyses, whereas age and BMI have very strong evidence and thereby are very significant factors in our analyses.

We also learn how much risk both age and BMI contribute. For every year increased in age, risk goes up by 5.7% and for every unit BMI increases, risk increases by 5%.

3.2 Testing our fit on a real patient

Now we've analysed our cox proportional hazards fit, it would be wise to fit a survival curve using some real-life data. We are going to take a future 50-year-old male patient with each of BMI = 20, 25, 30 and 35 and find the 30-day survival probability.

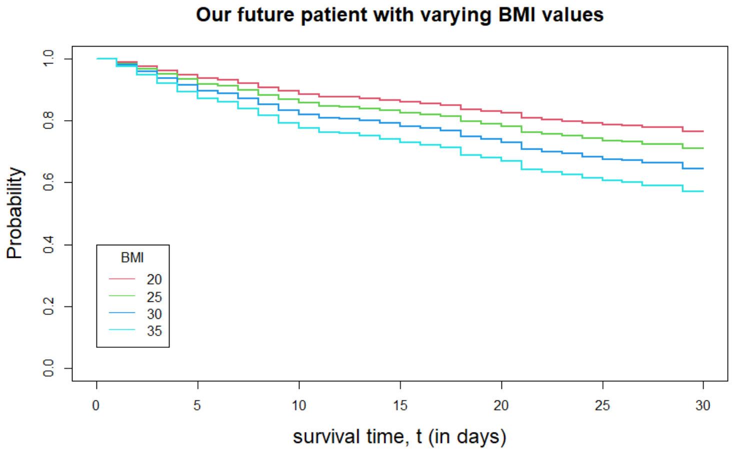

Figure 5: Survival curves for our future patient with varying BMI values

BMI	**S(15)**	**S(30)**
20	0.8593383	0.765
25	0.8238648	0.710
30	0.7806501	0.645
35	0.7287016	0.571

In the plot above we have our future patients survival curves plotted. The first thing we should note is that when we plot the survival curve, based on our cox proportional hazard model (where BMI takes on various values), none of our curves drop below 0.5 on the y-axis. Therefore we can't take a median survival time as we never reach that point in our given time interval. It can probably be assumed then that someone being healthy weight vs someone being morbidly obese doesn't have as great an impact on their survival time as we would have anticipated if the survival probabilities are both relatively high.

We deduced before that age is not a significant co variate so there's no reason to plot a set of curves where gender varies. However, we do know that age is significant (more significant than BMI according to the p-values).

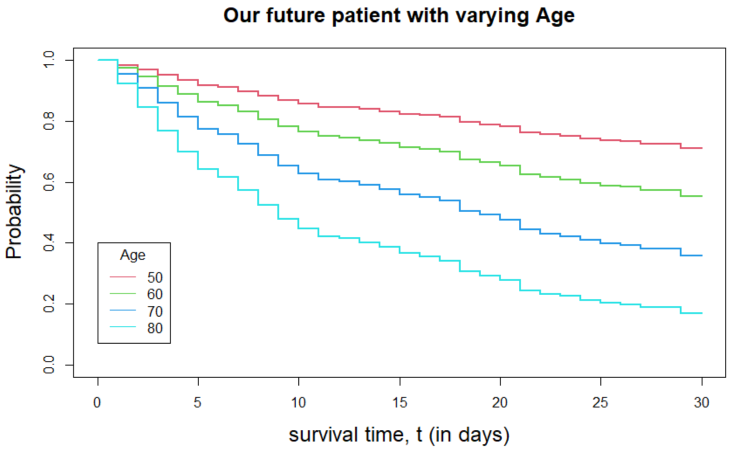

Figure 6: Survival curves for our future patient with varying Ages

Age	**Median**	**S(30)**
50	N/A	0.710
60	N/A	0.552
70	19	0.357
80	9	0.168

In the plot above we've kept gender and BMI constant whilst making Age the explanatory variable. Without needing to do much we can already see the huge difference in these plots. As age increases the survival time decreases aggressively. We now also can see how much greater risk there is in the elderly as we actually have probabilities low enough now that give us a median. Of course though it's difficult to compare age and BMI because their values have different interpretations in the context of what they are so we have to be careful with clearly stating that one is more significant than the other.

3.3 Adding another co variate to our model

For the sake of uncovering more information on our data we are going to fit the cox proportional hazards model again from this section, but this time we are going to add the co-variate ”Apache II Score” into our model and test to see whether it has any significance or not.

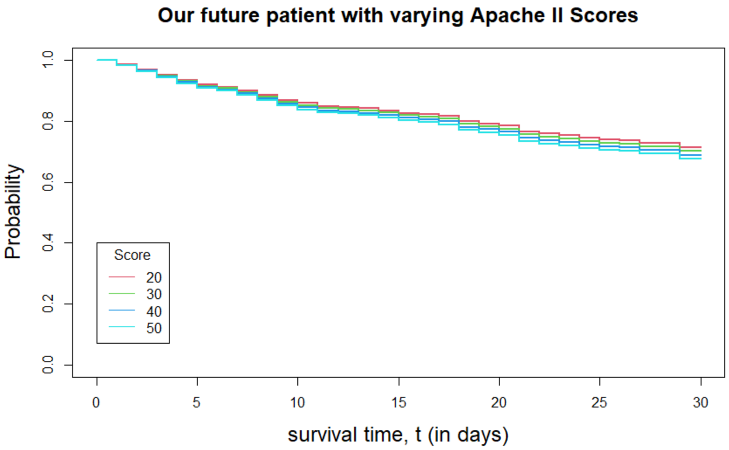

Figure 7: Survival curves for our future patient with varying Apache II Scores

We can see in figure 7 that, although the Apache II scores differ significantly, their survival times do not. Carrying out a summary on this model in R, we find that Apache II scores have a p-value of 0.462 (>>0.05). Therefore I think it's safe to conclude that this co-variate is not significant and won't add much in our final analysis.

4 A full analysis of the data

4.1 Our proposed significant covariates

through the tests we have completed in this report, we've been able to make some inferences about which covariates appear to have significance and which covariates do not. these things are not set in stone by any means but in order to aid our analysis we are going to attempt to rank them from least to most significant based off of their p-values relative to the tests we carried out on them.

Order	**Co-variate**	**Significant?**
Most important	Dependencies	yes
	Age	yes
	BMI	yes
	Invasive Ventilation	yes
	Comorbidities	yes
	Gender	no
	Apache II Score	no
Least important	Deprivation	unsure

4.2 Building and adjusting our models

We now want to build a model which best fits our data. We are therefore going to fit a full model and remove covariates one-by-one whilst measuring their AIC, C-index and R^2 values, using these values to draw conclusions about our models.

Model	**Terms**	**AIC**	**C-index**	R^2
1	All terms	2903.96	0.714	20.6
2	Model 1 - Apache	2902.42	0.713	20.6
3	Model 2 - Comor	2902.03	0.711	20.1
4	Model 3 - Female	2902.02	0.709	20.5
5	Model 4 - Depriv	2903.2	0.71	20.1
6	Model 5 - Invent	2904.49	0.712	18.1
7	Model 6 - Depend	2907.27	0.708	17.3
8	Age	2931.67	0.69	14.1

4.2.1 Conclusions about our model

Model 5 would be preferred. It may not have the lowest AIC value (although it's still relatively low) but it has a small number of explanatory variables (p=4) allows us to interpret the model much easier whilst maintaining a standard level of explained variation in the R^2 value too (the 4 explanatory variables are: Age, BMI, Dependencies and Invasive Ventilation). the C-indexes are all close together as well so this doesn't have much effect on our choice of model.

In conclusion, we have found the 4 covariates for our preferred model here to be the most influential on our data (as expected) but we also discovered that Comorbidities aren't quite as significant as we would have thought as well as Deprivation index being more significant than we thought, but there are always going to be unexpected outcomes like these.

4.3 Schoenfeld Residuals

Co-variate	**P-value**
Dependencies	0.22
Age	0.55
BMI	0.14
Invasive Ventilation	1.00
Comorbidities	0.32
Gender	0.71
Apache II Score	0.21
Deprivation	0.31
GLOBAL	0.11

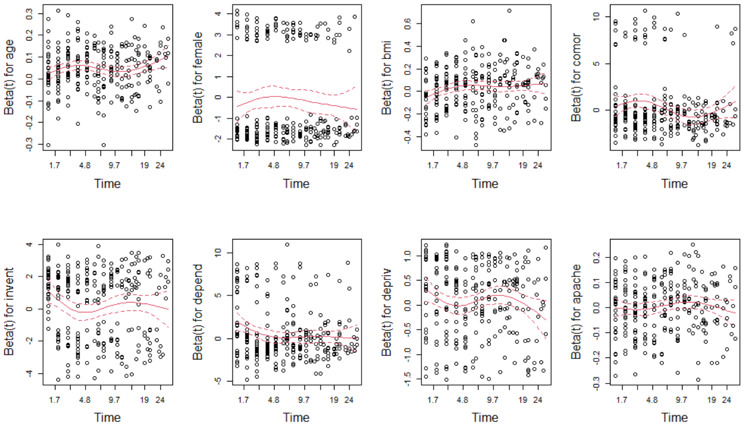

Figure 8: Schoenfeld residual plots for our co-variates

In the table above we have the p-values for our Schoenfeld residuals. As we can see, none of them have significant p-values. This suggests none of them have any trends and thereby none of them need to be investigated any further.

However, we see some kind of systematic trend in 3 of our plots (namely: Comorbidities, Deprivation and Invasive Ventilation) so we can assume that the proportional hazard assumption has been violated and thereby the hazard ratios for these 3 variables are changing over time.

4.3.1 Stratification

It may be wise then to stratify these covariates. We will only stratify the co-variate, ”Invasive Ventilation” because the other 2 covariates haven’t been selected to be part of our model.

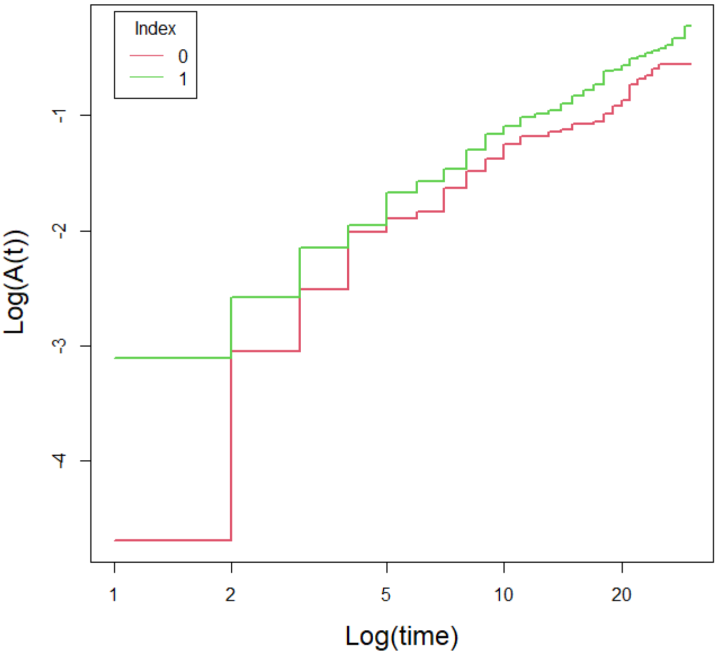

Figure 9: A stratified hazard curve for invasive ventilation

As you can see here though, the curves have a relatively similar shape so the proportionality assumption holds. Therefore there is no need to re-classify this co-variate into groups.

4.4 Conclusions

To conclude, we have come to discover that when fitting a model, we have 4 co-variates that we have found to be significant (Age, BMI, Dependencies and Invasive Ventilation). A co-variate such as comorbidities has significance but due to other restrictions it would not be wise to include it in our model (i.e. it doesn't add much in the way of explained variation).

There are also factors that just didn't appear to be significant enough at all (Gender, Apache II Score, Deprivation). Therefore we haven't included them in our final model as we don't believe they contribute enough to explain the model for what they are but the first 4 co-variates do.

We may also suggest that the sample itself is a little too short with regard to event time. We see a unusually high number of censoring at t=30, implying that our sample time is too short and didn't allow for enough events to occur before the end of the test. I think a time of 45 days may have been more appropriate for this data.

References

[1] *Apache II online score calculator*, Medscape URL = https://reference.medscape.com/calculator/12/apache-ii