Survival Analysis of Cancer Patients in the MIMIC-III Dataset

DATA 621 - Final Project | University of Calgary

Gohil Shilpa

April 28, 2020

Research Question

Assessing Cancer patient outcomes in MIMIC-III dataset using time spent in ICU as the period of observation, using a Cox Proportional Hazards Regression to determine whether cancer patients that undergo surgery have favorable outcomes compared to cancer patients that do not undergo surgery.

Introduction/Background

Cancer accounts for a significant portion of deaths around the world. In 2019, cancer was the second leading cause of death, following heart disease (World Health Organization). Previous research has shown the most common cancers worldwide are lung cancer (2.09 million cases), breast cancer (2.09 million cases) and colorectal cancer (1.80 million cases) while the most common causes of cancer death are colorectal cancer (862,000 deaths) and stomach cancer (783,000 deaths) (World Health Organization). Considerable variation and complexity exists in treating patients as cancer can affect any part of the body and is recognized as a large group of diseases with at least 65 recognized types of cancer (Zaorsky et al., 2016). The complex nature of this disease makes choices in cancer care particularly challenging.

The goal of this study is to understand differences in survivorship between different cancer patients. Therefore, we started by exploring whether there is a time-effect (duration of time spent in ICU) on overall survival of a cancer patient. The prognosis estimation is vital for the treatment process and dynamic prediction can allow registering early stages of unfavourable changes in patients' condition. Also important among cancer patients is early advance care planning conversations that lead to care that is concordant with patients' goals and wishes, particularly at the end of life.

In addition, we wanted to explore differences in survivorship for cancer patients that were admitted to the ICU as either 'urgent', 'emergency' or 'elective' type of admission . We also explore differences in survivorship between male and female cancer patients. This allows us to quantify and explore further why certain groups of cancer patients may fare better than others. Understanding these differences may help alleviate the personal, psychosocial and physical burden to the individual survivors.

Furthermore, we wanted to understand if there is a benefit or favourable outcome (efficacy) if a cancer patient undergoes surgery versus no surgery. Surgery is one of the major pillars of cancer care and control; it can be preventative (remove tissue that is likely to become cancer), diagnostic (biopsy), curative, supportive, palliative and reconstructive. Although surgical excision of primary or even metastatic tumours can save or extend life, it has long been acknowledged that the surgical insult itself may precipitate or accelerate tumor recurrence. Despite overwhelming evidence from experimental studies, clinical studies have not been as persuasive, and the concept is still subject to debate and the true impact it has on cancer patients remains unclear.

Therefore, we wanted to explore whether there is a time-effect (duration of time spent in ICU) on overall survival of a cancer patient.

Rationale/Objective

The present study was undertaken with the following objectives:

1. Determining whether there is a time spent in ICU effect on overall survival of cancer patients.
2. To understand the differences in survivorship between different cancer patients (males and females) and patients admitted as either elective, emergency or urgent.
3. To understand whether surgery such as cardiac surgery, neurological surgery, general surgery, thoracic surgery and vascular surgery improve overall survivorship for cancer patients.

Methods and Material

Study Design

This is a retrospective study describing part of the content of MIMIC-III dataset which contains information about a cohort of critically ill patients hospitalized from 2001 to 2012 at Beth Israel Deaconess Medical Centre (Boston, MA, USA). These patients were hospitalized due to diagnosis and complications arising from various types of cancers. Therefore, this study is exploratory in nature, with the goal of understanding prognostic factors of cancer patients' survival time. Information was derived from the electronic medical records of 46,476 unique critical care patients admitted to the intensive care unit. After completing a National Institutes of Health (NIH) web-based training courses (Protecting Human Research Participants), approval was obtained to download and use MIMIC-III data for the purposes of this study.

Study Population

MIMIC-III is available online and patient data was de-identified in a Health Insurance Portability and Accountability Act- compliant manner. We included as cancer patients that were admitted to ICU from 2001 to 1012. This included patients with various kinds of cancers; brain tumor, adrenal neoplasm, lung cancer, bladder cancer, multiple myeloma and many others (a total of 757 different cancer diagnosis).

Study variables

For our research purposes, out of 26 tables contained in the MIMIC-III data, we used 3 tables; ADMISSIONS, PATIENTS and SERVICES. Some information in these tables have been changed so that identities of patients can be protected. For example, the date of birth of a given patient from the PATIENTS table has been shifted to obscure their age and comply with HIPAA.

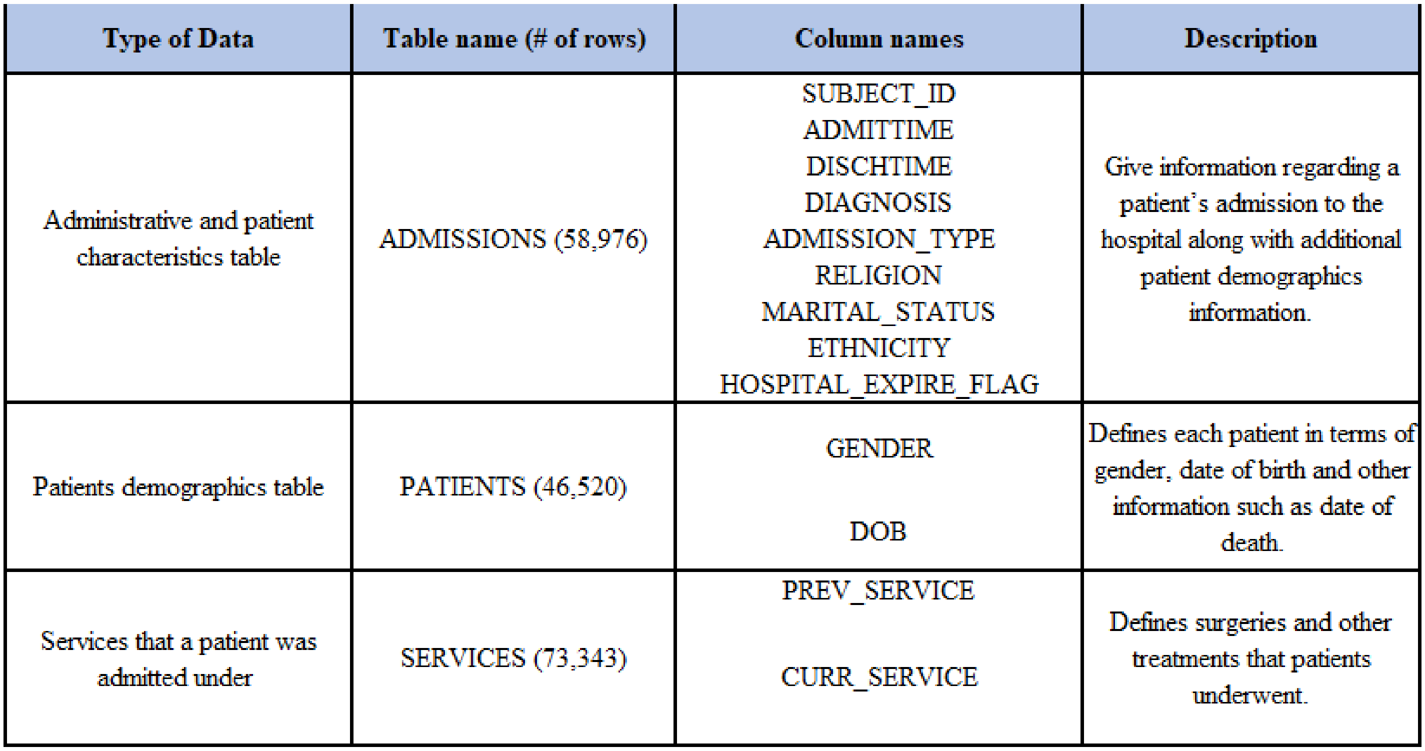
From the ADMISSIONS table, we included the columns SUBJECT_ID (unique identifier), ADMITTIME (time of admission), DISCHTIME (discharge time) and DIAGNOSIS in order to obtain information about all patients with cancer and to calculate the number of days spent in ICU by each cancer patient. For additional patient information we also included ADMISSION_TYPE (type of admission), RELIGION (type of religion), MARITAL_STATUS (marital status), ETHNICITY (ethnicity) and HOSPITAL_EXPIRE_FLAG (whether the patient is alive or dead).

From the PATIENT table, we included the columns GENDER (patients' sex) and DOB (date of birth) for additional patient demographic information and in order to calculate the age of each cancer patient.

From the SERVICES table, we included columns PREV_SERVICE and CURR_SERVICE in order to obtain information about surgeries that cancer patients may have had.

A note to be made here is, the time stamps for various variables, date of birth for all the patients in MIMIC-III data is modified in the tables, for privacy and confidentiality reason.

Table 1: An overview of the data tables and columns from MIMIC-III used in the study



Type of Data	Table name (# of rows)	Column names	Description
Administrative and patient characteristics table	ADMISSIONS (58,976)	SUBJECT_ID ADMITTIME DISCHTIME DIAGNOSIS ADMISSION_TYPE RELIGION MARITAL_STATUS ETHNICITY HOSPITAL_EXPIRE_FLAG	Give information regarding a patient's admission to the hospital along with additional patient demographics information.
Patients demographics table	PATIENTS (46,520)	GENDER DOB	Defines each patient in terms of gender, date of birth and other information such as date of death.
Services that a patient was admitted under	SERVICES (73,343)	PREV_SERVICE CURR_SERVICE	Defines surgeries and other treatments that patients underwent.

Data Processing

All data wrangling, cleaning and calculations were done in a Jupyter notebook using Pandas (Python) in the Anaconda environment and R Studio. The majority of the cleaning was done in Python (file attached seperately) to be able to work with large size files and to have various variables merged in one file seamlessly. Further bit of cleaning was completed in R (code attached in Appendix). From the ADMISSION table, patients with missing data (demographics, diagnosis) or patients with alternative diagnosis other than cancer were filtered out and excluded. There were a total of 750+ unique cancer diagnoses in our data.

We subsequently merged this data with the PATIENT table (on unique subject ID) and calculated each cancer patient's age and each patients' hospitalization stay by calculating DISCHTIME (discharge time) - ADMITTIME (admission time).

For surgeries performed on cancer patients, we merged the data from the SERVICES table and filtered all the data pertaining to cancer patients who had had either cardiac surgery (CSURG), neurological surgery (NSURG), general surgery (SURG), thoracic surgery (TSURG) and vascular surgery (VSURG). We excluded treatments such as plastic surgery (PSURG) and other non-surgical treatments such as neurological medical (NMED) and general service for internal medicine (MED).

Analysis

```
#loading the dataset
cancer = read.csv("Cancer_Surgery.csv", header = TRUE)
```

General Descriptive Statistics for the Cancer patient data.

1394 patients met the inclusion criteria and were included in the study. Among this cohort of patients, about 53% (727 patients) were male and 47% (667 patients) were female. The median age of these patients was 62 years and the overall average stay for all patients was approximately 13 days. A breakdown of patient average stay by type of admission is as follows:

- Patients admitted under urgent admission type: Average length of stay in ICU 25 days.
- Patients admitted under emergency admission type: Average length of stay in ICU 16 days.
- Patients admitted under elective admission type: Average length of stay in ICU is 10 days.

The formula applied to calculate the in-hospital mortality is:

> $$In\ hospital\ mortality = \frac{No\ of\ cancer\ deaths}{Total\ number\ of\ cancer\ patients}$$

The overall in-hospital mortality rate for all cancer patients in MIMIC-III for the period of 2001-2012 was 15.6%. The in-hospital mortality by type of admission can be broken down as follows:

- In-hospital mortality for cancer patients admitted under urgent admission: average mortality of 3.67%.
- In-hospital mortality for cancer patients admitted under emergency admission: average mortality of 79.91%.
- In-hospital mortality for cancer patients admitted under elective admission: average mortality of 16.51%.

Cancer ICU Admission rate for MIMICIII data:

$$Cancer\ Admission\ Rate = \frac{Total\ number\ of\ cancer\ cases}{(Total\ number\ of\ patients) \times (Total\ number\ of\ years)} = 0.0027\ person - years$$

$$Cancer\ Admission\ Rate = \frac{1394}{(46{,}520) \times (11\ years)} = 0.0027\ person - years$$

The rate of admission of cancer patients to ICU unit is 0.0027241 person-years or 2.724 per 1000 person-years.

Logistic Regression to determine whether there is an effect of time spent in ICU (days) on overall survival of cancer patients.

The question we would like to answer by logistic modeling is if there is an effect on overall patient mortality depending on time spent in ICU (in days) as our main exposure. Our response variable is whether the cancer patient is dead or alive (1 indicates death and 0 indicates survival upon discharge). As the outcome (response variable) is qualitative and only two outcomes are possible (binary), we opted for the logistic regression model to answer this question.

Therefore we use the logistic function:

$$E(y) = P(y = 1|X) = \frac{e^{\beta 0+\beta 1X}}{1 + e^{\beta 0+\beta 1X}}$$
$$where$$
$$y = 1 \; if \; cancer \; patient \; dies$$
$$y = 0 \; if \; cancer \; patient \; is \; alive \; and \; discharged$$

To fit the logistic model we use a method called *maximum likelihood estimation.*

Building the logistic model

Variables included in the study are as follows:

Stay_int = Number of days spent in ICU by cancer patients

Age = Age of cancer patient

Sex = Patient sex (male or female)

Type = Type of patient admission into ICU (elective, emergency or urgent)

We fit the full model with main exposure (time spent in ICU), age, sex and admission type:

```
#Full-model
full_logit_model <- glm(factor(Delta) ~ Stay_int + Age + Sex + Stay_int*Age + Stay_int*S
ex + factor(Type), data = cancer, family = "binomial")
summary(full_logit_model)
```

```
## 
## Call:
## glm(formula = factor(Delta) ~ Stay_int + Age + Sex + Stay_int * 
##     Age + Stay_int * Sex + factor(Type), family = "binomial", 
##     data = cancer)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.3324  -0.7398  -0.3083  -0.2834   2.5630  
## 
## Coefficients:
##                         Estimate Std. Error z value             Pr(>|z|)    
## (Intercept)           -3.3325780  0.2983917 -11.168 < 0.0000000000000002 ***
## Stay_int               0.0006298  0.0125568   0.050             0.959996    
## Age                    0.0033228  0.0031914   1.041             0.297805    
## SexM                  -0.1442607  0.2059803  -0.700             0.483702    
## factor(Type)EMERGENCY  1.9682764  0.1967051  10.006 < 0.0000000000000002 ***
## factor(Type)URGENT     1.6721499  0.4501670   3.715             0.000204 ***
## Stay_int:Age           0.0002224  0.0001838   1.211             0.226068    
## Stay_int:SexM          0.0021130  0.0078490   0.269             0.787776    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1208.6  on 1392  degrees of freedom
## Residual deviance: 1046.1  on 1385  degrees of freedom
## AIC: 1062.1
## 
## Number of Fisher Scoring iterations: 5
```

From the output of the model, we see that type of admission is significant however no other variables are significant (p-value greater than the default value of $\alpha = 0.05$). Therefore, we can go ahead and drop the sex variable.

```
#Dropping sex
logit_model2 <- glm(factor(Delta) ~ Stay_int + Age  + factor(Type) + Stay_int*Age, data
 = cancer, family = "binomial")
summary(logit_model2)
```

```
## 
## Call:
## glm(formula = factor(Delta) ~ Stay_int + Age + factor(Type) + 
##     Stay_int * Age, family = "binomial", data = cancer)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.3417  -0.7420  -0.3066  -0.2849   2.5540  
## 
## Coefficients:
##                         Estimate Std. Error z value             Pr(>|z|)    
## (Intercept)           -3.4069388  0.2803244 -12.154 < 0.0000000000000002 ***
## Stay_int               0.0018509  0.0117180   0.158             0.874494    
## Age                    0.0033405  0.0031930   1.046             0.295476    
## factor(Type)EMERGENCY  1.9699345  0.1965418  10.023 < 0.0000000000000002 ***
## factor(Type)URGENT     1.6757634  0.4493885   3.729             0.000192 ***
## Stay_int:Age           0.0002194  0.0001836   1.195             0.232072    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1208.6  on 1392  degrees of freedom
## Residual deviance: 1046.7  on 1387  degrees of freedom
## AIC: 1058.7
## 
## Number of Fisher Scoring iterations: 5
```

From the model above, we see that the interaction between number of days spent in ICU and patient’s age in not significant as p-value greater than the default value of $\alpha = 0.05$. Therefore, we can say that patient’s age is not an effect modifier for our main exposure (time spent in days in ICU for cancer patients).

```
#Dropping interaction
logit_model3 <- glm(factor(Delta) ~ Stay_int + Age  + factor(Type) , data = cancer, fami
ly = "binomial")
summary(logit_model3)
```

```
##
## Call:
## glm(formula = factor(Delta) ~ Stay_int + Age + factor(Type),
##     family = "binomial", data = cancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4548  -0.7436  -0.3104  -0.2843   2.5627
##
## Coefficients:
##                        Estimate Std. Error z value             Pr(>|z|)
## (Intercept)           -3.560092   0.247607 -14.378 < 0.0000000000000002 ***
## Stay_int               0.015020   0.003938   3.815             0.000136 ***
## Age                    0.006016   0.002212   2.720             0.006522 **
## factor(Type)EMERGENCY  1.957785   0.195656  10.006 < 0.0000000000000002 ***
## factor(Type)URGENT     1.659224   0.448437   3.700             0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1208.6  on 1392  degrees of freedom
## Residual deviance: 1048.1  on 1388  degrees of freedom
## AIC: 1058.1
##
## Number of Fisher Scoring iterations: 5
```

From the output above, we can see that all the variables are now significant including our main exposure (number of days spent in ICU), cancer patient's age, and type of admission (elective, emergency and urgent). Therefore, this is our final model.

We can do an anova to establish whether the non-interaction model is indeed a better model than the interaction model.

We can confirm that the non-interaction model is better than the interaction model:

H_0 : reduced model is true (model with no interaction term) i.e. all $\beta_i = 0$

$$\text{where i range for Age, Stay\_int, Type}$$

H_A : larger model is true (with interaction term) i.e. atleast one $\beta_i \neq 0$

$$\text{where i range for Age, Stay\_int, Type}$$

```
#Comparing interaction model with non-interaction model
anova(logit_model2, logit_model3, test = "Chisq")
```

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	1387	1046.687	NA	NA	NA
2	1388	1048.142	-1	-1.45515	0.2277034

2 rows

```
#likelihood ratio test
lrtest(logit_model2, logit_model3)
```

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	6	-523.3434	NA	NA	NA
2	5	-524.0710	-1	1.45515	0.2277034

2 rows

From the two outputs above, by using the Likelihood ratio test, we can see that the p-value = 0.2277 > $\alpha = 0.05$, for $\triangle G^2 = 1.4552$. Therefore, we fail to reject the null-hypothesis and hence, we can say that the main effects model is better than the interaction model. We should therefore choose the non-interaction model `(logit_model3)` as our best fit model. Since, the interaction is not significant between number of days spent in ICU and and age, we can say that there is no effect modification. Effect modification is when a variable that differentially (positively and/or negatively) modifies the observed effect of a risk factor on a disease status.

```
#Final model
mylogit <- glm(factor(Delta) ~ Stay_int + Age  + factor(Type) , data = cancer, family =
"binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = factor(Delta) ~ Stay_int + Age + factor(Type),
##     family = "binomial", data = cancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4548  -0.7436  -0.3104  -0.2843   2.5627
##
## Coefficients:
##                        Estimate Std. Error z value             Pr(>|z|)
## (Intercept)           -3.560092   0.247607 -14.378 < 0.0000000000000002 ***
## Stay_int               0.015020   0.003938   3.815             0.000136 ***
## Age                    0.006016   0.002212   2.720             0.006522 **
## factor(Type)EMERGENCY  1.957785   0.195656  10.006 < 0.0000000000000002 ***
## factor(Type)URGENT     1.659224   0.448437   3.700             0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1208.6  on 1392  degrees of freedom
## Residual deviance: 1048.1  on 1388  degrees of freedom
## AIC: 1058.1
##
## Number of Fisher Scoring iterations: 5
```

We can test the full model using Wald's test to check for the relationship between response and predictors

Setting up our hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_A : \ at\ least\ one\ \beta_i \neq 0$
$where\ i : \ 1 = \ ICU\ Stay, \ 2 = \ Age, \ 3 = \ Emergency\ Admission, \ 4 = \ Urgent\ Admission$

```
wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 3)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 7.4, df = 1, P(> X2) = 0.0065
```

From the output, the Wald χ^2 = 7.4 with p-value = 0.0065. This indicates that the probability of patient mortality depends on at least one predictor at $alpha = 0.05$.

We can go ahead and calculate 95% the confidence intervals for the coefficients of our variables.

Confidence interval for the logit model is,

```
confint(mylogit)
```

```
## Waiting for profiling to be done...
```

```
##                                2.5 %      97.5 %
## (Intercept)            -4.061267577 -3.08669116
## Stay_int                0.007133195  0.02267273
## Age                     0.001489630  0.01024475
## factor(Type)EMERGENCY   1.586113149  2.35520266
## factor(Type)URGENT      0.725850066  2.50402950
```

From the confidence interval, we see that none of the 95% confidence intervals capture 0 and hence we can say that the variables are significant.

```
coef(mylogit)
```

```
##           (Intercept)              Stay_int                   Age
##          -3.560092409           0.015020433           0.006015989
## factor(Type)EMERGENCY    factor(Type)URGENT
##           1.957784856           1.659224184
```

To calculate the antilog for β_i:

```
sum.coef<-summary(mylogit)$coef
est<-exp(sum.coef[,1])
print(est)
```

```
##           (Intercept)              Stay_int                   Age
##             0.0284362             1.0151338             1.0060341
## factor(Type)EMERGENCY    factor(Type)URGENT
##             7.0836184             5.2552322
```

We can also check for confounding of variables in the model. A variable that can cause or prevent the outcome of interest and the effects cannot be distinguished from those of other factors being studied is said to be confounding. This happens when the association between the exposure and the outcome is obscured by a third variable or factor that:

a. Is associated with the exposure

b. Is associated or is independent factor for the outcome

Since our main exposure is time spent in ICU by cancer patients (in days), we can check to see if age is a confounder.

```
#Checking for confounding with main exposure and age
mylogit_age <- glm(factor(Delta) ~ Stay_int  + factor(Type) , data = cancer, family = "b
inomial")
summary(mylogit_age)
```

```
##
## Call:
## glm(formula = factor(Delta) ~ Stay_int + factor(Type), family = "binomial",
##     data = cancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4736  -0.7513  -0.3110  -0.2980   2.5101
##
## Coefficients:
##                        Estimate Std. Error z value             Pr(>|z|)
## (Intercept)           -3.135376   0.181280 -17.296 < 0.0000000000000002 ***
## Stay_int               0.014472   0.003929   3.683             0.000230 ***
## factor(Type)EMERGENCY  1.927801   0.193901   9.942 < 0.0000000000000002 ***
## factor(Type)URGENT     1.619701   0.447625   3.618             0.000296 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1208.6  on 1392  degrees of freedom
## Residual deviance: 1054.7  on 1389  degrees of freedom
## AIC: 1062.7
##
## Number of Fisher Scoring iterations: 5
```

We find that after dropping age from our model the main exposure $\beta_{Days\ spent\ in\ ICU}$ changes less than 10% (changes by 6.67%). Therefore, we can conclude that age is not confounding with our main exposure, days spent in ICU by cancer patients.

Now, let us see if the type of admission is a confounder:

```
#Checking for confounding with main exposure and type of admission
mylogit_type <- glm(factor(Delta) ~ Stay_int  + Age , data = cancer, family = "binomial"
)
summary(mylogit_type)
```

```
## 
## Call:
## glm(formula = factor(Delta) ~ Stay_int + Age, family = "binomial", 
##     data = cancer)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7967  -0.5707  -0.5383  -0.5101   2.1010  
## 
## Coefficients:
##              Estimate Std. Error z value             Pr(>|z|)    
## (Intercept) -2.255967   0.167523 -13.467 < 0.0000000000000002 ***
## Stay_int     0.019718   0.003807   5.179          0.000000223 ***
## Age          0.004258   0.002003   2.126               0.0335 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1208.6  on 1392  degrees of freedom
## Residual deviance: 1178.9  on 1390  degrees of freedom
## AIC: 1184.9
## 
## Number of Fisher Scoring iterations: 4
```

From the output above, we do see that days spent in ICU has a confounding effect with type of admission (whether elective, urgent or emergency). When admission variable is removed from the model, the main exposure $\beta_{Days\ spent\ in\ ICU}$ changes by more than 10% (changes by 26.7%). From this we can conclude that type of admission by cancer patients has a confounding effect with days spent in ICU.

Intrepretations for all the coefficients in our model.

$\widehat{beta_{Stayint}} = 0.015020433$. This indicates that an increase in days spent in ICU by a cancer patient is associated with an increase in the probability of patient mortality. To be more precise, a one-unit (day) increase in cancer patient stay at ICU is associated with an increase in the log odds of patient mortality by 0.0150 units.

$exp^{\beta_{Stayint}} = e^{0.015020433} = 1.015$. Therefore, for every additional day spent in ICU, we estimate the odds ratio of mortality to be multiplied by about 1.015 i.e there is an increase of 1.5% `[=(1-1.015)*100%]` odds of cancer patient mortality.

$\widehat{beta_{Age}} = 0.006015989$. This indicates that an increase in a cancer patients age is associated with an increase in the probability of patient mortality. To be more precise, a one unit increase in age (1 year) for a cancer patient is associated with increase in the log odds of patient mortality by 0.006 units.

$exp^{\beta_{Age}} = e^{0.006015989} = 1.0060341$. Therefore, for every increase in age by a year, we estimate the odds ratio of mortality to be multiplied by about 1.006 i.e there is an increase of 0.6% `[=(1.0060-1)*100%]` odds of patient mortality.

$\widehat{beta_{Emergency}} = 1.957784856$. This indicates that a cancer patient admitted due to emergency admission is associated with an increase in patient mortality as compared to a patient admitted due to elective admission. To be more precise, a cancer patient admitted as an emergency admission is associated with an increase in log odds of patient mortality by 1.958 units.

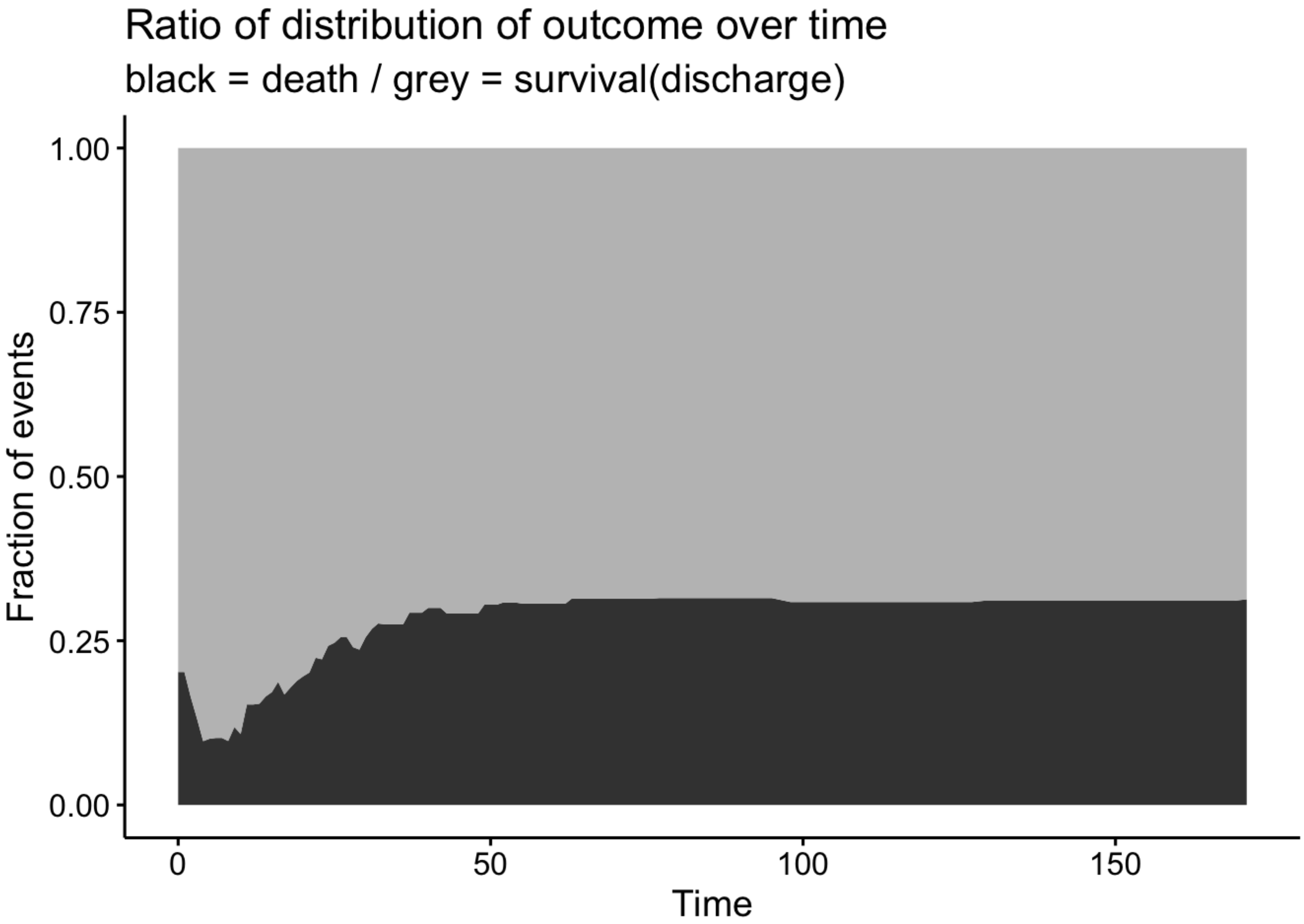
$exp^{\beta_{Emergency}} = e^{1.957784856} = 7.083$. Therefore, if a cancer patient is admitted as an emergency admission, we estimate the odds ratio of mortality to be multiplied by about 7.083 i.e there is an increase of 608.3% `[(7.083-1)*100]` odds of patient mortality as compared to a patient admitted under elective admission.

$\widehat{beta_{Urgent}} = 1.659224184$. This indicates that a cancer patient admitted due to urgency admission is associated with an increase in patient mortality as compared to a patient admitted due to elective admission. To be more precise, a cancer patient admitted as an emergency admission is associated with an increase in log odds of patient mortality by 1.660 units.

$exp^{\beta_{Urgent}} = e^{1.659224184} = 5.255$. Therefore, if a cancer patient is admitted as an urgent admission, we estimate the odds ratio of mortality to be multiplied by about 7.083 i.e there is an increase of 425.5% `[(5.255-1)*100]` odds of patient mortality as compared to a patient admitted under elective admission.

Graphical Distribution of the effect of time spent in ICU (in days) with cancer patient mortality.

```
surv <- Surv(cancer$Stay_int, cancer$Delta)
ggsurvevents(surv, main="Ratio of distribution of outcome over time", submain="black = d
eath / grey = survival(discharge)")
```



The function ggsurvevents() [in survminer] calculates and plots the distribution for patient death events (both status = 0 and status = 1). As we can see from the output above, as cancer patients spend more days in ICU (time variable in x-axis), the ratio of deaths (delta = 1) to survivals (delta = 0) increases as shown by increased grey as compared to black area in the graph above.

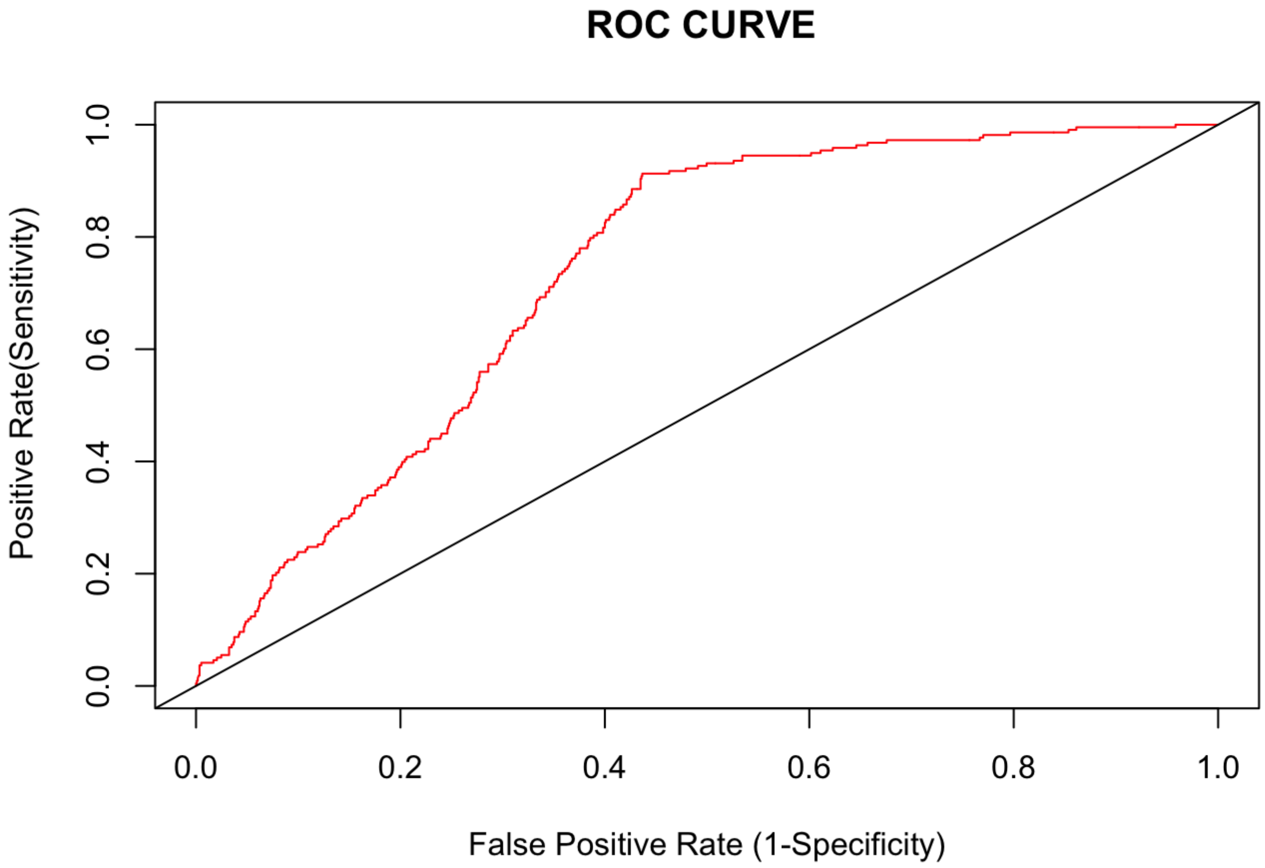
Evaluating the logistic regression model

1. Deviance - The deviance is a measure of goodness of fit of a generalized linear model. The null deviance shows how well the response variable is predicted by the model that only includes the intercept. The residual deviance indicates how well the response is predicted by the model with independent variables. In our model, we have a value of 1208.6 on 1392 degrees of freedom. Including the independent variables (days spent in ICU, age, and type of admission), decreased the deviance to 1048.1 on 1388 degrees of freedom, a significant reduction in deviance. The residual deviance has reduced by 160.5 points with a loss of 4 degrees of freedom.

2. AIC (Akaike Information Criteria) - The AIC provides a method for assessing the quality of the model through comparison or related models and is based on deviance. Our final model has an AIC value of 1058.1 and that is slightly less than the interaction model which has an AIC value of 1058.7. In general, model with a smaller AIC value is a better model.

3. ROC curve (Receiver Operating Characteristic) - The ROC curve is a plot to represent the true positive rate (sensitivity), the probability of predicting a real positive will be a positive, against false positive rate (1-specificity), the probability of predicting a real negative will be positive. A good model should be high on sensitivity and low on 1-specificity. It's a rule that predicts most true positives will be a positive and few true negatives will be a positive.

```
#The ROC curve for our final model
cancer$Delta = factor(cancer$Delta)
prob1=predict(mylogit,type=c("response"))
pred<-prediction(prob1,cancer$Delta)
perf<-performance(pred,measure ="tpr",x.measure="fpr")
plot(perf,col=2,main="ROC CURVE ", xlab="False Positive Rate (1-Specificity)",ylab="True
     Positive Rate(Sensitivity)")
abline(0,1)
```

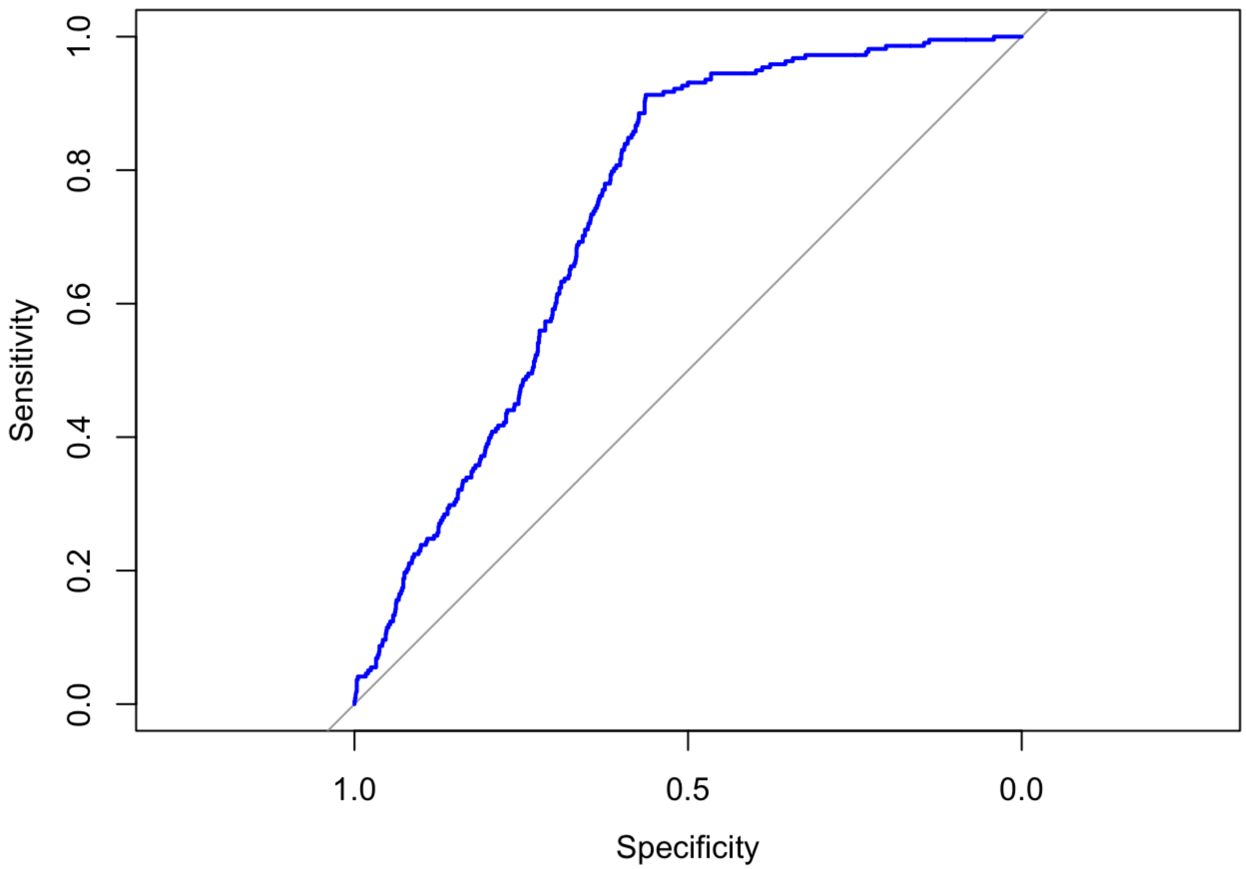


```
roc2<-roc(cancer$Delta,prob1)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(roc2, col="blue")
```



```
auc(roc2)
```

```
## Area under the curve: 0.7384
```

We find that the AUC (Area under the curve) for our model is 0.738.

Hence we define our final logistic model as:

$$\widehat{Cancer\ Patient\ Mortality} = -3.560 + 0.0150 Days\ in\ ICU + 0.006 Age + 1.958\ Emergency + 1.659\ Urgent$$

The logistic regression model is:

$$\hat{\pi} = \frac{e^{-3.560+0.0150 Days\ in\ ICU+0.006 Age+1.958 Emergency+1.659 Urgent}}{1+e^{-3.560+0.0150 Days\ in\ ICU+0.006 Age+1.958 Emergency+1.659 Urgent}}$$

$$\hat{\pi} = \begin{cases} = \frac{e^{-3.560+0.0150+0.006}}{1+e^{-3.560+0.0150+0.06}} = 0.0282 & \text{if Cancer Patient Admitted as Elective Admission} \\ = \frac{e^{-3.560+0.0150+0.006+1.958}}{1+e^{-3.560+0.0150+0.006+1.958}} = 0.1706 & \text{if Cancer Patient Admitted as Emergency Admission} \\ = \frac{e^{-3.560+0.0150+0.006+1.958+1.659}}{1+e^{-3.560+0.0150+0.006+1.958+1.659}} = 0.0237 & \text{if Cancer Patient Admitted as Urgent Admission} \end{cases}$$

Example prediction 1:

$$P(Cancer\ Patient\ Mortality|Age\ 70|EmergencyAdmission) = \frac{e^{-3.560+0.0150+0.006(70)+1.958}}{1+e^{-3.560+0.0150+0.006(70)+1.958}} = 0.2375$$

Example prediction 2:

$$P(Cancer\ Patient\ Mortality|Age\ 55|ElectiveAdmission) = \frac{e^{-3.560+0.0150+0.006}}{1+e^{-3.560+0.0150+0.06}} = 0.1744$$

The predicted mortality probability for a 70 year old cancer patient admitted to emergency at the Beth Israel Deaconess Medical Centre between 2001 and 2012 is 0.23% if all else is kept constant. The predicted mortality for a 55 year old patient also admitted to emergency at Beth Israel Deaconess Medical Centre between 2001 and 2012 has a predicted mortality probability of 0.1744%, if all else is kept constant. This indicates that a 70 year old patient in emergency type admission has a higher mortality probability than a 55 year old patient in elective admission type (a cancer patient that is both younger and is admitted under a different admission has a lower probability).

Survival Analysis

Kaplan Meier Curves helps understand the differences in survivorship between different cancer patients (males and females) and patients admitted as either elective, emergency or urgent.

The two most important measures in cancer studies include:

1. The time to death
2. The relapse-free survival time, which corresponds to the time between response to treatment and recurrence of the disease (also know as disease-free survival time and event-free survival time)

In this study, we are interested in exploring the time to event (patient mortality) for different cancer patient cohorts (male and female cancer patients) and cancer patients admitted to ICU under either elective, emergency and urgent admission.

Two related probabilities are used to describe survival data: the survival probability and the hazard probability. The survival probability, also known as the survivor function $S(t)$, is the probability that an individual survives from the time origin (e.g admission to ICU) to a specified future time t. The hazard, denoted by $h(t)$, is the probability that an individual who is under observation at a time t has an outcome of the event at that time. Thus, in contrast to the survivor function, which focuses on not having an event, the hazard function focuses on the event occurring.

The Kaplan-Meier (KM) method is a non-parametric method used to estimate the survival probability from observed survival times. The survival probability at time ti, $S(t_i)$, is calculated as follows:

$$S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i})$$

$$\mathit{Where}\ S(t_{i-1}) = \mathit{the\ probability\ of\ being\ alive\ at}\ (t_{i-1})$$

$$n_i = \mathit{the\ number\ of\ patients\ alive\ just\ before}\ t_i$$

$$d_i = \mathit{the\ number\ of\ events\ at}\ t_i$$

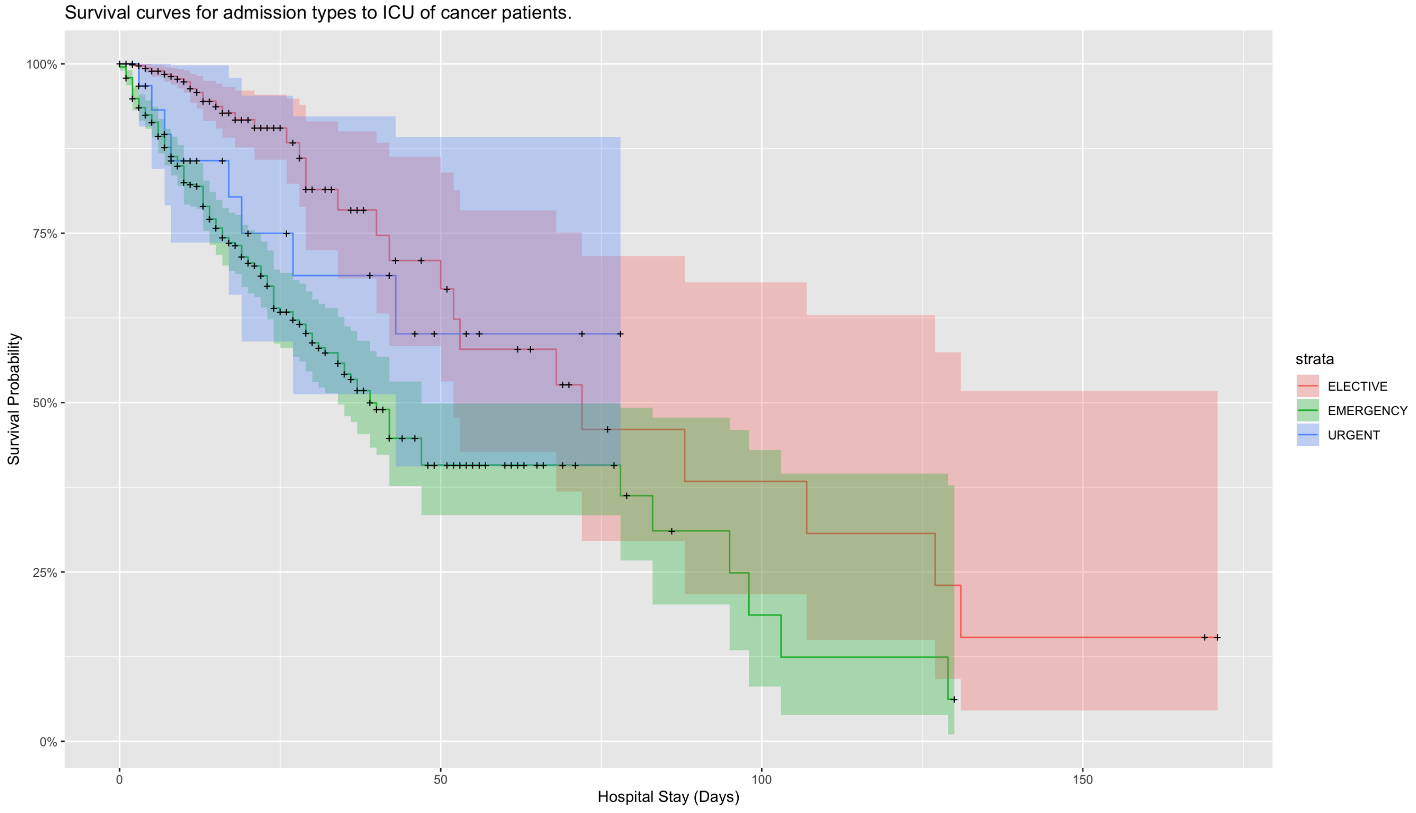
$$t_0,\ S(0) = 1$$

The estimated probability $(S(t))$ is a step function that changes in value only at the time of each event. Thus, we use the Kaplan-Meier survival curve, a plot of KM survival probability against time, as it provides a useful summary of the data that can be used to estimate measures such as mean survival time.

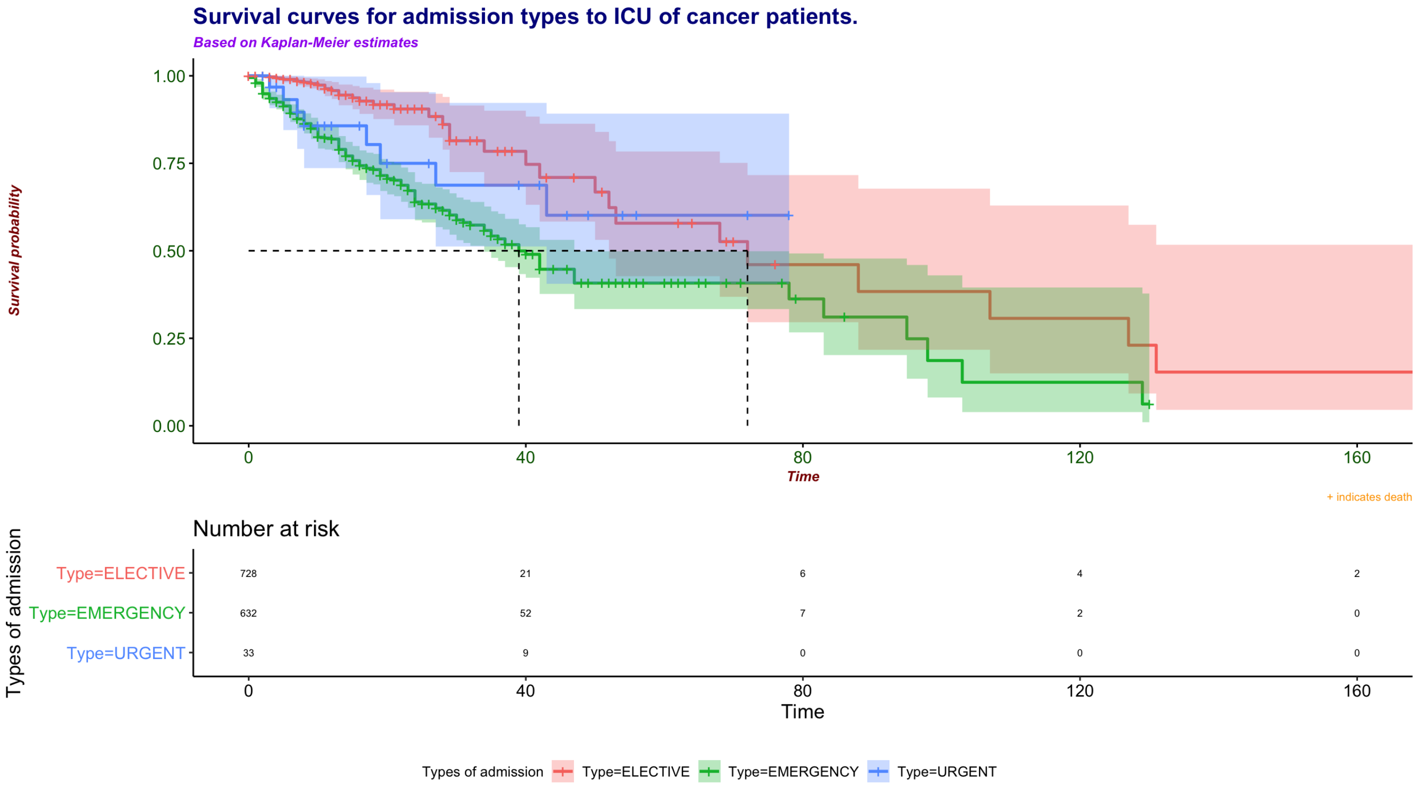
Therefore, we compute the survival analyses of cancer patients in MIMIC-III admitted under either elective, emergency and urgent admission. We subsequently use a Kaplan-Meier plot to summarize and visualize the results of the survival analysis.

Kaplan-Meier Curve to show Survival Analysis of cancer patients admitted as either elective, emergency or urgent admission.

```
cancer$Type = factor(cancer$Type)
km_trt_fit <- survfit(Surv(Stay_int, Delta) ~ Type, data=cancer)
autoplot(km_trt_fit,  xlab = 'Hospital Stay (Days)', ylab = 'Survival Probability', main
="Survival curves for admission types to ICU of cancer patients.", font.main=15)
```



```
#using survminer.
survminer::ggsurvplot(km_trt_fit, size=1,title = "Survival curves for admission types to
ICU of cancer patients.", subtitle = "Based on Kaplan-Meier estimates", caption = "+ ind
icates death", legend="bottom",
           legend.title="Types of admission",
   font.title = c(16, "bold", "darkblue"),
   font.subtitle = c(10, "bold.italic", "purple"),
   font.caption = c(8, "plain", "orange"),
   font.x = c(10, "bold.italic", "darkred"),
   font.y = c(10, "bold.italic", "darkred"),
   font.tickslab = c(12, "plain", "darkgreen"),
                       conf.int=T,
                       censor=T,
                       surv.median.line = "hv",
                       risk.table = TRUE,
                       risk.table.height=.3,
                      fontsize=2.5)
```



Types of admission	0	40	80	120	160
Type=ELECTIVE	728	21	6	4	2
Type=EMERGENCY	632	52	7	2	0
Type=URGENT	33	9	0	0	0

From the Kaplan Meier curves above, we see some interesting results. For cancer patients that were admitted under emergency admission many more patients experiences the event (death) as compared to patients admitted under urgent or elective treatment. We also see that at around 40 days, emergency cancer patients have a 50% survival probability as compared to cancer patients admitted under elective surgery, who have a 50% survival probability of about 75 days. For cancer patients admitted under urgent admission, there are fewer patients that experience the event however, it is also important to note that there are fewer patients in that cohort to begin with. We see that the confidence interval are quite wide for cancer patients admitted under urgent admission, thus giving us a clue that the study contains very few participants. Interestingly, we also see widening of confidence intervals for elective cancer patients after around 30 days and emergency cancer patients at around 80 days, thus also indicating that the study contains very few participants at the respective time as the participants in the study experience the event of interest (death) and the number of participants decrease as the study goes on.

Log-rank test.

Filtering data for elective and emergency patients in order to do a log rank test.

```
elect <- cancer[cancer$Type == "ELECTIVE",]
emerg <- cancer[cancer$Type == "EMERGENCY",]

#Concatenating data
combined = rbind(elect, emerg)
```

Performing the log-rank test.

The log rank test is the primary tool for the comparison of the survival estimates of two or more groups. The log rank test compares the entire survival experience between groups and can be thought of as a test of whether the survival curves are identical (overlapping) or not. It is designed particularly to detect a difference between survival curves which result when mortality in one group is higher than the corresponding one in the second group and the ratio of these is constant over time. The log rank test is closely linked to the chi-square test statistic and compares observed to expected numbers of events at each point over the follow-up period. In our study, we have a cross over between the elective and urgent survival curves. Therefore, it would not be appropriate to conduct a log rank test to detect whether there is a difference between cancer patients admitted under elective admission and those admitted under urgent admission. However, since we do not have a cross-over between elective and emergency admitted cancer patients, it would be beneficial and appropriate to filter the data and perform a log-rank test for elective and emergency admitted cancer groups in order to test whether there is a difference between the two groups in terms of their respective survival curves.

H_0 : The two survival curves are identical (the survival estimate between emergency cancer patients and elective cancer patients is identical) (or $S_{1t} = S_{2t}$)

H_A : The two survival curves are not identical (the survival estimate between emergency cancer patients and elective cancer patients is not identical) (or $S_{1t} \neq S_{2t}$, at any time t) ($\alpha = 0.5$)

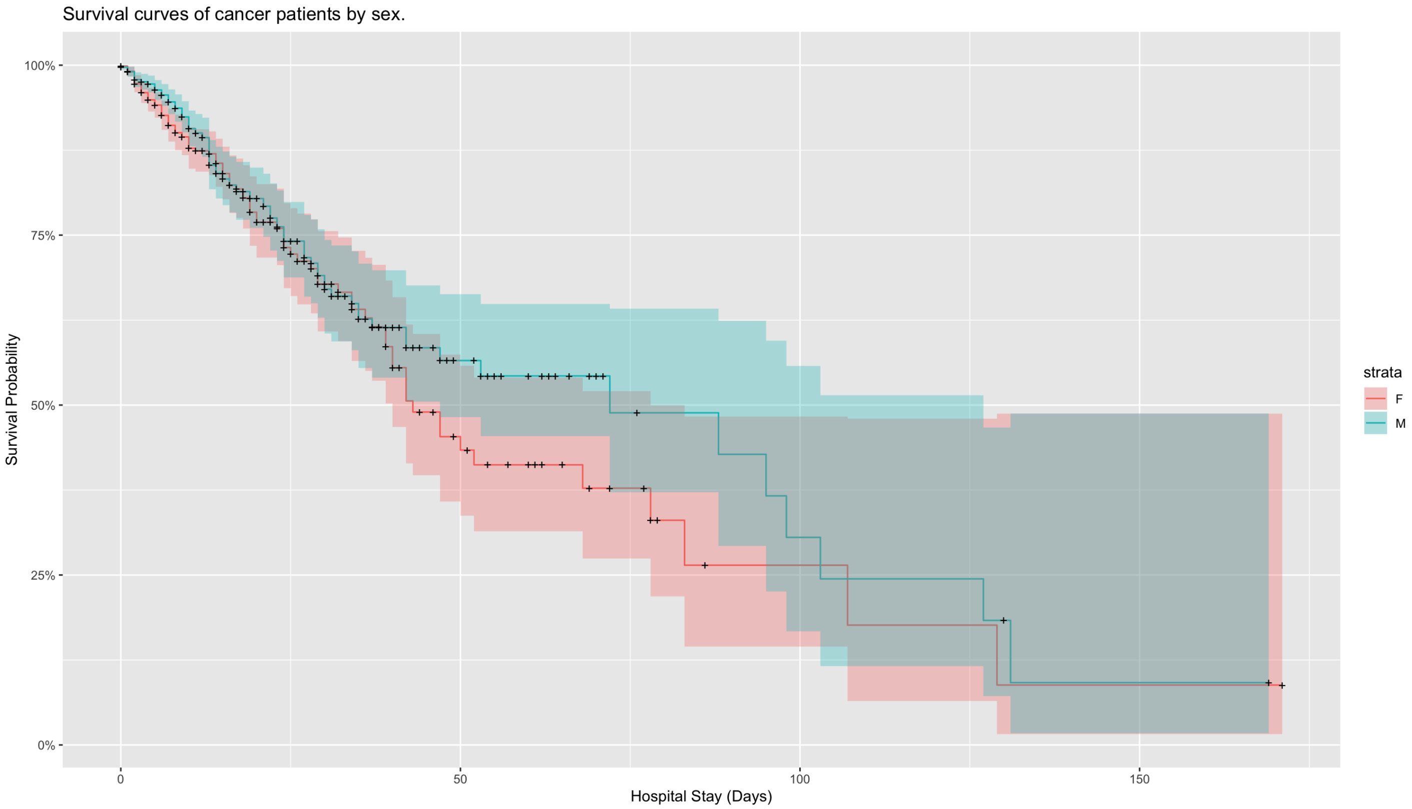
```
#Log rank test
survdiff(Surv(Stay_int,Delta)~Type,data=combined)
```

```
## Call:
## survdiff(formula = Surv(Stay_int, Delta) ~ Type, data = combined)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## Type=ELECTIVE  728       36     87.9      30.6      55.9
## Type=EMERGENCY 632      174    122.1      22.0      55.9
##
##  Chisq= 55.9  on 1 degrees of freedom, p= 0.00000000000007
```

From the output above, the value of chi-square statistic is 55.9 with 1 degrees of freedom and p-value is therefore 0.00000000000007 and hence we would reject the null-hypothesis in favour of the alternative hypothesis as 0.00000000000007 < 0.05 (default value of alpha). Thus, we would reject the null hypothesis that the survival estimate between emergency cancer patients and elective cancer patients is identical and therefore we could say that the survival estimates between the elective and emergency patients is not identical and hence elective cancer patients can be said to fare better than emergency cancer patients.

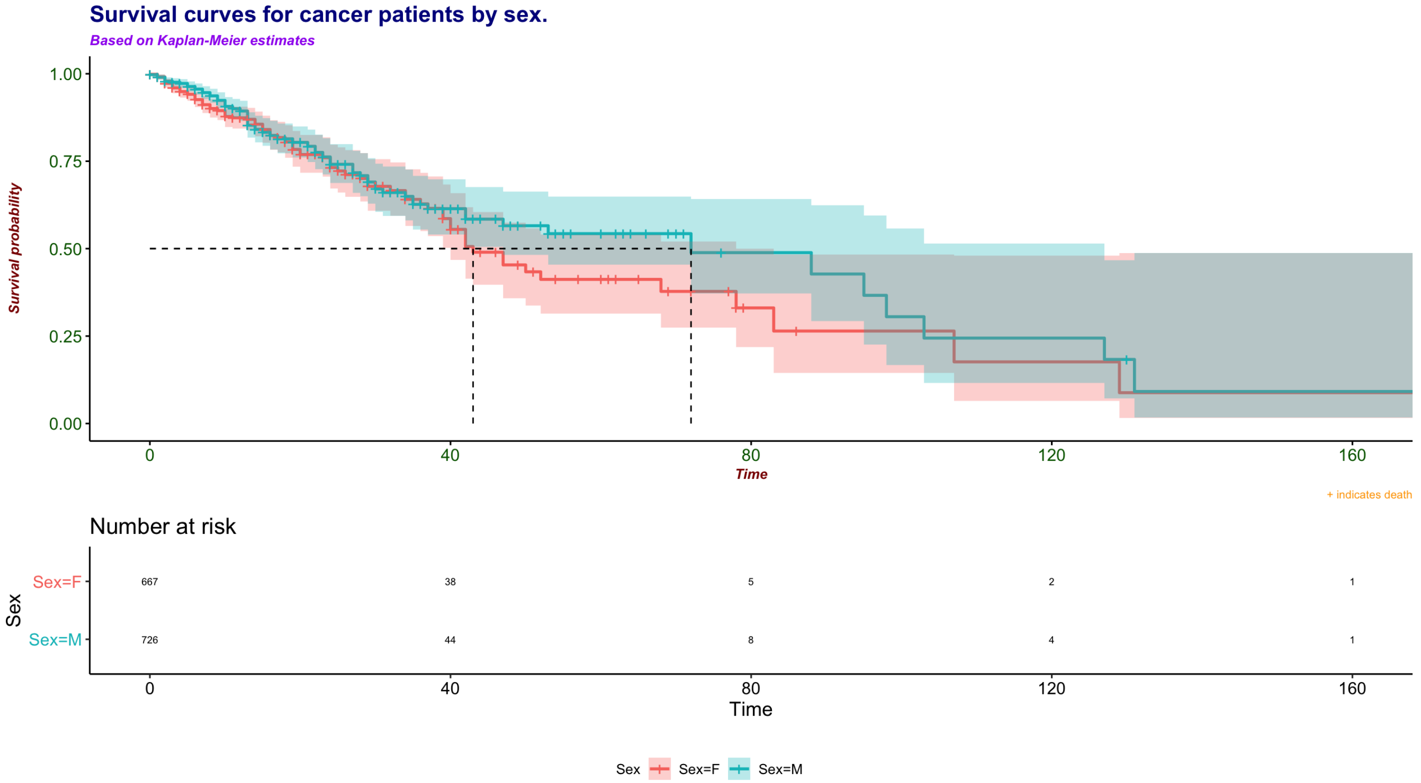
Kaplan-Meier Curve to show Survival Analysis of Male and Female cancer patients.

```
cancer$Type = factor(cancer$Type)
sur_fit <- survfit(Surv(Stay_int, Delta) ~ Sex, data=cancer)
autoplot(sur_fit,  xlab = 'Hospital Stay (Days)', ylab = 'Survival Probability', main="S
urvival curves of cancer patients by sex.")
```



```
#using survminer.

survminer::ggsurvplot(sur_fit, size=1,
                    title = "Survival curves for cancer patients by sex.", subtitle = "B
ased on Kaplan-Meier estimates", caption = "+ indicates death", legend="bottom",
           legend.title="Sex",
   font.title = c(16, "bold", "darkblue"),
   font.subtitle = c(10, "bold.italic", "purple"),
   font.caption = c(8, "plain", "orange"),
   font.x = c(10, "bold.italic", "darkred"),
   font.y = c(10, "bold.italic", "darkred"),
   font.tickslab = c(12, "plain", "darkgreen"),
                       conf.int=T,
                       censor=T,
                       surv.median.line = "hv",
                       risk.table = TRUE,
                       risk.table.height=.3,
                      fontsize=2.5)
```



From the Kaplan Meier curves above, we see that both many male and female patients experience the event (death) starting almost from day 1 and continuing till day 40 whereby we see that the confidence intervals start to widen for both groups thus indicating that the study contains fewer patients after the 40 day mark as the participants in the study experience the event of interest and the number of participants decrease as the study goes on. However, no conclusive results can be obtained from this Kaplan Meier model as we see that there is much crossing over between the two survivor curves and hence it would not be accurate to perform a log-rank test.

Cox Proportional Hazards Regression for Survival Analysis.

The Cox proportional-hazards model (Cox, 1972) is essentially a regression model commonly used statistical method in medical research for investigating and studying the association between the survival time of patients in a certain category and one or more explanatory variables. Cox proportional hazards regression analysis works for both quantitative as well as for categorical variables in our predictions. In clinical research and studies, there are many situations, where several known quantities (known as covariates), potentially affect patient prognosis. The cox proportional-hazards model is one of the most important methods used for modelling survival analysis data.

For our analyses in regards to the survival of patients with some form of cancer, we will be looking at the effect of different types of surgery (if any) that a patient might undergo, while in the ICU over the time of stay in the hospital.

The Cox model is expressed by the *hazard function* denoted by $h(t)$. Briefly, the hazard function can be interpreted as the risk of death or survival at time t. It can be estimated as follows:

$$h(t) = h_0(t) \times \exp(\beta_1 x_1 \; + \; \beta_2 x_2 \; + \; \ldots\ldots + \beta_m x_m)$$

Where:

- t represents the survival time.
- $h(t)$ is the hazard function determined by a set of m covariates $(x_1, x_2, \ldots., x_m)$.
- the coefficients $(\beta_1, \beta_2, \ldots, \beta_m)$ measure the impact (i.e., the effect size) of respective covariates.
- the term h_0 is called the baseline hazard and $h_0(t)$ is the baseline hazard function or a nuisance function. It corresponds to the value of the hazard if all the x_s are equal to zero (the quantity exp(0) equals 1).
- $exp(\beta_m)$ are called hazard ratios (HR).

The $h(t)$ is a hazard function which varies over time as t changes.

The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables x_s, with the baseline hazard being an 'intercept' term that varies with time.

Generally, a value of β_m greater than 0, or $exp(\beta_m)$ greater than 1, indicates that as the value of the m^{th} covariate increases, the event hazard increases and the time length of survival decreases. Thus, in short it tells us that the covariate with $HR > 1$ is positively associated with the event and negatively associated with the length of survival. When $HR = 0$ there is no effect of that covariate on the event.

> Note: In cancer studies:
>
> - A covariate with hazard ratio > 1 (i.e.: b > 0) is called bad prognostic factor
> - A covariate with hazard ratio < 1 (i.e.: b < 0) is called good prognostic factor

After describing the general Cox proportional hazard regression model, we will start looking at building our own model. For this we will work with the hypothesis we want to test, set a basic model and look for effects of different predictors on the model. Important notion to consider are whether there are any *a priori* hypotheses to be tested and whether a multiple variable model is even needed. Since, there is often a considerable amount of variability between subjects in a medical study, finding a parsimonious model that is relevant to the outcome of interest is very important.

Testing the hypothesis for distribution of time to death are the same for different surgery types experienced by the cancer patients.

With this, we will now define our basic model. Our significance level (α) is set at 0.05. We will be relying on the Wald Statistic values (z) in our summary outputs to determine the significance of the variable in the model. The Wald statistic evaluates, whether the beta (β) coefficient of a given variable is statistically significantly different from 0.

The covariates that we are interested in studying that may potentially affect cancer patient prognosis from a clinical study standpoint are:

Surgery (main-exposure for the study) = Type of surgery patient surgery (cardiac surgery, neurological surgery, general surgery, thoracic surgery and vascular surgery).

Sex = Male and female patients in MIMIC-III data

Age = Patient’s age

Type = Type of ICU admission (either elective, emergency or urgent). This is mainly chosen by us because generally to study such patient cohort, we have some form of severity index. Since, the MIMIC data was missing it, we felt that this variable might be a close option to test severity of cancer patients after their surgery.

Note: Hospital (ICU) stay time in our analysis over which the outcome is measured is set to begin after the patient undergoes the surgery. (Details discussed in limitations later)

We first set our reference level to “no-surgery” as surgery is our main exposure and we want to primarily compare patient prognosis for cancer patients that have undergone surgery to cancer patients that have not undergone any surgery.

Building a univariate cox regression with only our main exposure (type of surgery).

We now build a univariate cox model with our main exposure as type of surgery.

```
#Univariate model with main exposure (type of surgery)
options(scipen = 999)
cox_main_exposure <- coxph(Surv(Stay_int, Delta) ~ factor(Surgery) , data =  surgery, ti
es = "breslow")
summary(cox_main_exposure)
```

```
## Call:
## coxph(formula = Surv(Stay_int, Delta) ~ factor(Surgery), data = surgery,
##     ties = "breslow")
##
##   n= 1393, number of events= 218
##
##                                        coef      exp(coef)        se(coef)
## factor(Surgery)Cardiac        -1.1167114751    0.3273545400    0.5839084425
## factor(Surgery)Neurological   -0.8905491347    0.4104303093    0.2739543318
## factor(Surgery)General        -0.8885564870    0.4112489677    0.2162891917
## factor(Surgery)Thorasic       -0.2097747737    0.8107668314    0.2416026519
## factor(Surgery)Vascular      -14.3773943334    0.0000005701 1415.4983664792
##                                    z  Pr(>|z|)
## factor(Surgery)Cardiac      -1.912   0.05582 .
## factor(Surgery)Neurological -3.251   0.00115 **
## factor(Surgery)General      -4.108 0.0000399 ***
## factor(Surgery)Thorasic     -0.868   0.38525
## factor(Surgery)Vascular     -0.010   0.99190
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                exp(coef)  exp(-coef) lower .95 upper .95
## factor(Surgery)Cardiac      0.3273545400       3.055    0.1042    1.0281
## factor(Surgery)Neurological 0.4104303093       2.436    0.2399    0.7021
## factor(Surgery)General      0.4112489677       2.432    0.2692    0.6284
## factor(Surgery)Thorasic     0.8107668314       1.233    0.5049    1.3018
## factor(Surgery)Vascular     0.0000005701 1753973.482    0.0000       Inf
##
## Concordance= 0.635  (se = 0.02 )
## Likelihood ratio test= 32.82  on 5 df,   p=0.000004
## Wald test            = 27.13  on 5 df,   p=0.00005
## Score (logrank) test = 29.61  on 5 df,   p=0.00002
```

From the output above, we see that the p-value for all three overall tests (likelihood ratio test, Wald and logrank) are significant, indicating that the model is significant. These tests evaluate the omnibus null hypothesis that all the β_s are 0. From our model, the omnibus null hypothesis is soundly rejected as p-value is less than the default value of $\alpha = 0.05$. The Wald statistic evaluates, whether the β coefficient of our variable (type of surgery) is statistically significantly different from 0. Therefore, from the output, we can conclude that the two types of surgeries are significant; neurological and general surgery.

We now proceed to a multivariate cox regression analysis to see how our other variables jointly impact survival. Therefore, we will add age, sex and type of admission to our main exposure (type of surgery). We also add interaction term to see if there is an effect modification between type of surgery that cancer patient undergoes and age.

Multivariate Cox Regression Analysis including interaction between age and type of surgery to see how factors jointly impact survival.

```
#Multivariate cox regression model with interaction
options(scipen = 9)
cox_full <- coxph(Surv(Stay_int, Delta) ~ factor(Surgery) + factor(Sex) + factor(Type) +
Age + factor(Surgery)*Age, data =  surgery, ties = "breslow")
summary(cox_full)
```

```
## Call:
## coxph(formula = Surv(Stay_int, Delta) ~ factor(Surgery) + factor(Sex) + 
##     factor(Type) + Age + factor(Surgery) * Age, data = surgery, 
##     ties = "breslow")
## 
##   n= 1393, number of events= 218 
## 
##                                           coef      exp(coef)       se(coef)
## factor(Surgery)Cardiac             0.398836836    1.490090469    2.418300728
## factor(Surgery)Neurological       -0.135193798    0.873546604    0.856362786
## factor(Surgery)General            -0.539179875    0.583226375    0.353822533
## factor(Surgery)Thorasic            1.083687140    2.955557037    0.929336099
## factor(Surgery)Vascular          -13.154228988    0.000001937 7853.728624555
## factor(Sex)M                      -0.126840332    0.880874308    0.137053974
## factor(Type)EMERGENCY              1.239687912    3.454535178    0.196983425
## factor(Type)URGENT                 0.762588873    2.143819118    0.404751312
## Age                                0.008344693    1.008379607    0.002367977
## factor(Surgery)Cardiac:Age        -0.020765497    0.979448621    0.040354744
## factor(Surgery)Neurological:Age   -0.008002615    0.992029320    0.013188106
## factor(Surgery)General:Age        -0.001443995    0.998557048    0.003451099
## factor(Surgery)Thorasic:Age       -0.013122822    0.986962907    0.014710575
## factor(Surgery)Vascular:Age       -0.021074100    0.979146407  112.461819418
##                                       z       Pr(>|z|)    
## factor(Surgery)Cardiac            0.165       0.869003    
## factor(Surgery)Neurological      -0.158       0.874559    
## factor(Surgery)General           -1.524       0.127541    
## factor(Surgery)Thorasic           1.166       0.243579    
## factor(Surgery)Vascular          -0.002       0.998664    
## factor(Sex)M                     -0.925       0.354718    
## factor(Type)EMERGENCY             6.293 0.000000000311 ***
## factor(Type)URGENT                1.884       0.059552 .  
## Age                               3.524       0.000425 ***
## factor(Surgery)Cardiac:Age       -0.515       0.606851    
## factor(Surgery)Neurological:Age -0.607       0.543980    
## factor(Surgery)General:Age       -0.418       0.675643    
## factor(Surgery)Thorasic:Age      -0.892       0.372357    
## factor(Surgery)Vascular:Age       0.000       0.999850    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##                                    exp(coef)  exp(-coef) lower .95 upper .95
## factor(Surgery)Cardiac            1.490090469      0.6711 1.302e-02 1.705e+02
## factor(Surgery)Neurological       0.873546604      1.1448 1.631e-01 4.680e+00
## factor(Surgery)General            0.583226375      1.7146 2.915e-01 1.167e+00
## factor(Surgery)Thorasic           2.955557037      0.3383 4.782e-01 1.827e+01
## factor(Surgery)Vascular           0.000001937 516189.3777 0.000e+00       Inf
## factor(Sex)M                      0.880874308      1.1352 6.734e-01 1.152e+00
## factor(Type)EMERGENCY             3.454535178      0.2895 2.348e+00 5.082e+00
## factor(Type)URGENT                2.143819118      0.4665 9.698e-01 4.739e+00
## Age                               1.008379607      0.9917 1.004e+00 1.013e+00
## factor(Surgery)Cardiac:Age        0.979448621      1.0210 9.050e-01 1.060e+00
## factor(Surgery)Neurological:Age 0.992029320      1.0080 9.667e-01 1.018e+00
## factor(Surgery)General:Age        0.998557048      1.0014 9.918e-01 1.005e+00
## factor(Surgery)Thorasic:Age       0.986962907      1.0132 9.589e-01 1.016e+00
## factor(Surgery)Vascular:Age     0.979146407      1.0213 1.833e-96 5.230e+95
##
## Concordance= 0.73  (se = 0.02 )
## Likelihood ratio test= 95.47  on 14 df,   p=3e-14
## Wald test            = 48.45  on 14 df,   p=0.00001
## Score (logrank) test = 93.14  on 14 df,   p=1e-13
```

From the output above, we see that interaction between type of surgery and age is not significant as p-value > 0.05 (default value of alpha). Therefore, we drop the interaction from our model.

Multivariate Cox Regression Analysis without interaction.

```
#Multivariate cox regression model without interaction
cox_full1 <- coxph(Surv(Stay_int, Delta) ~ factor(Surgery) + factor(Sex) + factor(Type)
 + Age, data =  surgery, ties = "breslow")
summary(cox_full1)
```

```
## Call:
## coxph(formula = Surv(Stay_int, Delta) ~ factor(Surgery) + factor(Sex) + 
##     factor(Type) + Age, data = surgery, ties = "breslow")
## 
##   n= 1393, number of events= 218 
## 
##                                      coef       exp(coef)        se(coef)
## factor(Surgery)Cardiac        -0.9381499592    0.3913511817    0.5870909547
## factor(Surgery)Neurological   -0.6405075047    0.5270248885    0.2745636403
## factor(Surgery)General        -0.6269081808    0.5342410268    0.2247355982
## factor(Surgery)Thorasic        0.2608331462    1.2980110693    0.2488289433
## factor(Surgery)Vascular      -14.2627409445    0.0000006394 1415.0868156861
## factor(Sex)M                  -0.1266176026    0.8810705273    0.1369980288
## factor(Type)EMERGENCY          1.2633666550    3.5373103669    0.1974484765
## factor(Type)URGENT             0.7836438013    2.1894356167    0.4049236340
## Age                            0.0068663694    1.0068899970    0.0017222029
##                                   z       Pr(>|z|)    
## factor(Surgery)Cardiac      -1.598        0.11005    
## factor(Surgery)Neurological -2.333        0.01966 *  
## factor(Surgery)General      -2.790        0.00528 ** 
## factor(Surgery)Thorasic      1.048        0.29453    
## factor(Surgery)Vascular     -0.010        0.99196    
## factor(Sex)M                -0.924        0.35537    
## factor(Type)EMERGENCY        6.398 0.000000000157 ***
## factor(Type)URGENT           1.935        0.05295 .  
## Age                          3.987 0.000066922764 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##                             exp(coef)   exp(-coef) lower .95 upper .95
## factor(Surgery)Cardiac      0.3913511817       2.5552    0.1238    1.2368
## factor(Surgery)Neurological 0.5270248885       1.8974    0.3077    0.9027
## factor(Surgery)General      0.5342410268       1.8718    0.3439    0.8299
## factor(Surgery)Thorasic     1.2980110693       0.7704    0.7970    2.1139
## factor(Surgery)Vascular     0.0000006394 1563974.5765    0.0000       Inf
## factor(Sex)M                0.8810705273       1.1350    0.6736    1.1525
## factor(Type)EMERGENCY       3.5373103669       0.2827    2.4022    5.2088
## factor(Type)URGENT          2.1894356167       0.4567    0.9901    4.8418
## Age                         1.0068899970       0.9932    1.0035    1.0103
## 
## Concordance= 0.729  (se = 0.02 )
## Likelihood ratio test= 93.12  on 9 df,   p=4e-16
## Wald test            = 69.01  on 9 df,   p=0.00000000002
## Score (logrank) test = 84.62  on 9 df,   p=2e-14
```

From the output, we see that after dropping interaction term, neurological and general surgery remain significant however if a patient was admitted under emergency type admission and the patients age now becomes significant as p-value < 0.05 (default value of alpha). The p-value of a patient admitted as emergency admission is 0.000000000157, with a hazard ratio of exp(1.2633666550) = 3.537, indicating a strong relationship between cancer patients admitted as emergency admission and increased risk of mortality. The p-value of a cancer patients age is 0.000066922764, with a hazard ratio of exp(0.0068663694) = 1.0068, indicating a strong relationship between cancer patients age and increased risk of mortality.

We can perform an anova to compare the models with interaction and without interaction. Testing the hypothesis.

H_0 : All the $\beta_{interactions}$ are insignificant (i.e.$\beta_{interactions} = 0$).

H_A : Atleast one of the $\beta_{interactions}$ are significant (i.e.$\beta_{interactions} \neq 0$).

```
#Anova to compare model with interaction and without
anova(cox_full1, cox_full)
```

	loglik <dbl>	Chisq <dbl>	Df <int>	P(>|Chi|) <dbl>
1	-1246.965	NA	NA	NA
2	-1245.790	2.350168	5	0.7988712

2 rows

From the output above, we see p-value 0.7989 > default value of alpha at 0.05. Therefore, we accept the null-hypothesis and we can say that there is evidence to suggest that the model without interaction is a better model.

We now drop sex from our model as it was found to be insignificant in the previous model.

Multivariate Cox Regression Analysis without sex variable.

```
#Multivariate cox regression model - dropping sex
cox_full2 <- coxph(Surv(Stay_int, Delta) ~ factor(Surgery)  + factor(Type) + Age, data =
surgery, ties = "breslow")
summary(cox_full2)
```

```
## Call:
## coxph(formula = Surv(Stay_int, Delta) ~ factor(Surgery) + factor(Type) +
##     Age, data = surgery, ties = "breslow")
##
##   n= 1393, number of events= 218
##
##                                       coef      exp(coef)       se(coef)
## factor(Surgery)Cardiac        -0.9764909176    0.3766304099    0.5856643977
## factor(Surgery)Neurological   -0.6550603543    0.5194107130    0.2740079544
## factor(Surgery)General        -0.6236737390    0.5359717958    0.2246347883
## factor(Surgery)Thorasic        0.2531329280    1.2880544840    0.2488445024
## factor(Surgery)Vascular      -14.2196387993    0.0000006676 1404.4910358378
## factor(Type)EMERGENCY          1.2726022667    3.5701309175    0.1974159623
## factor(Type)URGENT             0.8111544641    2.2505046144    0.4037899827
## Age                            0.0068831196    1.0069068627    0.0017227570
##                                   z       Pr(>|z|)
## factor(Surgery)Cardiac      -1.667         0.0955 .
## factor(Surgery)Neurological -2.391         0.0168 *
## factor(Surgery)General      -2.776         0.0055 **
## factor(Surgery)Thorasic      1.017         0.3090
## factor(Surgery)Vascular     -0.010         0.9919
## factor(Type)EMERGENCY        6.446 0.000000000115 ***
## factor(Type)URGENT           2.009         0.0446 *
## Age                          3.995 0.000064582526 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                exp(coef)   exp(-coef) lower .95 upper .95
## factor(Surgery)Cardiac      0.3766304099       2.6551    0.1195    1.1870
## factor(Surgery)Neurological 0.5194107130       1.9253    0.3036    0.8887
## factor(Surgery)General      0.5359717958       1.8658    0.3451    0.8324
## factor(Surgery)Thorasic     1.2880544840       0.7764    0.7909    2.0977
## factor(Surgery)Vascular     0.0000006676 1497996.0396    0.0000       Inf
## factor(Type)EMERGENCY       3.5701309175       0.2801    2.4246    5.2568
## factor(Type)URGENT          2.2505046144       0.4443    1.0199    4.9658
## Age                         1.0069068627       0.9931    1.0035    1.0103
##
## Concordance= 0.729  (se = 0.019 )
## Likelihood ratio test= 92.26  on 8 df,   p=<2e-16
## Wald test            = 68.06  on 8 df,   p=0.00000000001
## Score (logrank) test = 83.52  on 8 df,   p=1e-14
```

From the output above, we can see that neurological and general surgery remain significant and patient's age and emergency admission remain significant and urgent emergency now also becomes significant. We can subsequently check to see if dropping the sex variable is a better model than the model with sex included.

H_0 : β_{sex} is insignificant (i.e.$\beta_{sex} = 0$).

H_A : β_{sex} is significant (i.e.$\beta_{sex} \neq 0$).

```
#checking to see if model with sex or without sex is better
anova(cox_full2,cox_full1)
```

	loglik	Chisq	Df	P(>|Chi|)
	<dbl>	<dbl>	<int>	<dbl>
1	-1247.392	*NA*	*NA*	*NA*
2	-1246.965	0.854135	1	0.3553852

2 rows

From the output above, we see that p-value 0.3554 > 0.05 (default value of alpha) and hence we accept the null hypothesis and we say that there is evidence to suggest that the model without sex is a better model. Therefore, we proceed with the model without sex.

We also check if there is an interaction between type of surgery and type of admission:

Checking for interaction between type of surgery and type of admission multivariate cox regression.

```
#Multivariate cox regression model - interaction between surgery and type of admission
cox_full3 <- coxph(Surv(Stay_int, Delta) ~ factor(Surgery)  + factor(Type) + Age + (fact
or(Surgery)*factor(Type)), data =  surgery, ties = "breslow")
summary(cox_full3)
```

```
## Call:
## coxph(formula = Surv(Stay_int, Delta) ~ factor(Surgery) + factor(Type) + 
##     Age + (factor(Surgery) * factor(Type)), data = surgery, ties = "breslow")
## 
##   n= 1393, number of events= 218 
## 
##                                                              coef
## factor(Surgery)Cardiac                             -0.6181982069
## factor(Surgery)Neurological                         0.0393413514
## factor(Surgery)General                             -0.4782792735
## factor(Surgery)Thorasic                             0.0351229581
## factor(Surgery)Vascular                           -14.4070107324
## factor(Type)EMERGENCY                               1.3817286204
## factor(Type)URGENT                                  0.6600895036
## Age                                                 0.0069363029
## factor(Surgery)Cardiac:factor(Type)EMERGENCY       -0.3555507806
## factor(Surgery)Neurological:factor(Type)EMERGENCY  -1.0062829287
## factor(Surgery)General:factor(Type)EMERGENCY       -0.3010056049
## factor(Surgery)Thorasic:factor(Type)EMERGENCY       0.3777752949
## factor(Surgery)Vascular:factor(Type)EMERGENCY      -1.4683856569
## factor(Surgery)Cardiac:factor(Type)URGENT         -14.2781761465
## factor(Surgery)Neurological:factor(Type)URGENT     -0.3721206829
## factor(Surgery)General:factor(Type)URGENT           1.5161795094
## factor(Surgery)Thorasic:factor(Type)URGENT                    NA
## factor(Surgery)Vascular:factor(Type)URGENT                    NA
##                                                         exp(coef)
## factor(Surgery)Cardiac                              0.5389145759
## factor(Surgery)Neurological                         1.0401254713
## factor(Surgery)General                              0.6198490654
## factor(Surgery)Thorasic                             1.0357470545
## factor(Surgery)Vascular                             0.0000005535
## factor(Type)EMERGENCY                               3.9817786658
## factor(Type)URGENT                                  1.9349655130
## Age                                                 1.0069604148
## factor(Surgery)Cardiac:factor(Type)EMERGENCY        0.7007873568
## factor(Surgery)Neurological:factor(Type)EMERGENCY   0.3655753267
## factor(Surgery)General:factor(Type)EMERGENCY        0.7400736247
## factor(Surgery)Thorasic:factor(Type)EMERGENCY       1.4590350535
## factor(Surgery)Vascular:factor(Type)EMERGENCY       0.2302969636
## factor(Surgery)Cardiac:factor(Type)URGENT           0.0000006296
## factor(Surgery)Neurological:factor(Type)URGENT      0.6892710543
## factor(Surgery)General:factor(Type)URGENT           4.5547903779
## factor(Surgery)Thorasic:factor(Type)URGENT                    NA
## factor(Surgery)Vascular:factor(Type)URGENT                    NA
##                                                          se(coef)      z
## factor(Surgery)Cardiac                               1.0493870982 -0.589
## factor(Surgery)Neurological                          0.4980876498  0.079
## factor(Surgery)General                               0.4354183302 -1.098
## factor(Surgery)Thorasic                              0.4959751095  0.071
## factor(Surgery)Vascular                           3741.0889030693 -0.004
## factor(Type)EMERGENCY                                0.2909886794  4.748
## factor(Type)URGENT                                   0.5294474379  1.247
## Age                                                  0.0017490779  3.966
## factor(Surgery)Cardiac:factor(Type)EMERGENCY         1.2685335204 -0.280
## factor(Surgery)Neurological:factor(Type)EMERGENCY    0.6147536445 -1.637
## factor(Surgery)General:factor(Type)EMERGENCY         0.5168672556 -0.582
## factor(Surgery)Thorasic:factor(Type)EMERGENCY        0.5679900768  0.665
## factor(Surgery)Vascular:factor(Type)EMERGENCY     5152.5685978600  0.000
## factor(Surgery)Cardiac:factor(Type)URGENT         2311.2798611703 -0.006
## factor(Surgery)Neurological:factor(Type)URGENT       1.2039444966 -0.309
## factor(Surgery)General:factor(Type)URGENT            0.9440800326  1.606
## factor(Surgery)Thorasic:factor(Type)URGENT           0.0000000000     NA
## factor(Surgery)Vascular:factor(Type)URGENT           0.0000000000     NA
##                                                     Pr(>|z|)
## factor(Surgery)Cardiac                                 0.556
## factor(Surgery)Neurological                            0.937
## factor(Surgery)General                                 0.272
## factor(Surgery)Thorasic                                0.944
## factor(Surgery)Vascular                                0.997
## factor(Type)EMERGENCY                             0.00000205 ***
## factor(Type)URGENT                                     0.212
## Age                                               0.00007318 ***
## factor(Surgery)Cardiac:factor(Type)EMERGENCY           0.779
## factor(Surgery)Neurological:factor(Type)EMERGENCY      0.102
## factor(Surgery)General:factor(Type)EMERGENCY           0.560
## factor(Surgery)Thorasic:factor(Type)EMERGENCY          0.506
## factor(Surgery)Vascular:factor(Type)EMERGENCY          1.000
## factor(Surgery)Cardiac:factor(Type)URGENT              0.995
## factor(Surgery)Neurological:factor(Type)URGENT         0.757
## factor(Surgery)General:factor(Type)URGENT              0.108
## factor(Surgery)Thorasic:factor(Type)URGENT                NA
## factor(Surgery)Vascular:factor(Type)URGENT                NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                      exp(coef)   exp(-coef)
## factor(Surgery)Cardiac                            0.5389145759       1.8556
## factor(Surgery)Neurological                       1.0401254713       0.9614
## factor(Surgery)General                            0.6198490654       1.6133
## factor(Surgery)Thorasic                           1.0357470545       0.9655
## factor(Surgery)Vascular                           0.0000005535 1806696.7438
## factor(Type)EMERGENCY                             3.9817786658       0.2511
## factor(Type)URGENT                                1.9349655130       0.5168
## Age                                               1.0069604148       0.9931
## factor(Surgery)Cardiac:factor(Type)EMERGENCY      0.7007873568       1.4270
## factor(Surgery)Neurological:factor(Type)EMERGENCY 0.3655753267       2.7354
## factor(Surgery)General:factor(Type)EMERGENCY      0.7400736247       1.3512
## factor(Surgery)Thorasic:factor(Type)EMERGENCY     1.4590350535       0.6854
## factor(Surgery)Vascular:factor(Type)EMERGENCY     0.2302969636       4.3422
## factor(Surgery)Cardiac:factor(Type)URGENT         0.0000006296 1588302.1071
## factor(Surgery)Neurological:factor(Type)URGENT    0.6892710543       1.4508
## factor(Surgery)General:factor(Type)URGENT         4.5547903779       0.2195
## factor(Surgery)Thorasic:factor(Type)URGENT                  NA           NA
## factor(Surgery)Vascular:factor(Type)URGENT                  NA           NA
##                                                   lower .95 upper .95
## factor(Surgery)Cardiac                              0.06891     4.215
## factor(Surgery)Neurological                         0.39184     2.761
## factor(Surgery)General                              0.26403     1.455
## factor(Surgery)Thorasic                                 0.39181     2.738
## factor(Surgery)Vascular                                 0.00000       Inf
## factor(Type)EMERGENCY                                   2.25106     7.043
## factor(Type)URGENT                                      0.68550     5.462
## Age                                                     1.00351     1.010
## factor(Surgery)Cardiac:factor(Type)EMERGENCY            0.05832     8.421
## factor(Surgery)Neurological:factor(Type)EMERGENCY       0.10957     1.220
## factor(Surgery)General:factor(Type)EMERGENCY            0.26873     2.038
## factor(Surgery)Thorasic:factor(Type)EMERGENCY           0.47928     4.442
## factor(Surgery)Vascular:factor(Type)EMERGENCY           0.00000       Inf
## factor(Surgery)Cardiac:factor(Type)URGENT               0.00000       Inf
## factor(Surgery)Neurological:factor(Type)URGENT          0.06510     7.298
## factor(Surgery)General:factor(Type)URGENT               0.71592    28.978
## factor(Surgery)Thorasic:factor(Type)URGENT                   NA        NA
## factor(Surgery)Vascular:factor(Type)URGENT                   NA        NA
##
## Concordance= 0.733  (se = 0.018 )
## Likelihood ratio test= 99.86  on 16 df,   p=4e-14
## Wald test            = 77.77  on 16 df,   p=0.0000000004
## Score (logrank) test = 98.67  on 16 df,   p=6e-14
```

From the output above, we see that non of the interactions are significant as p-values are more than the default value of alpha at 0.05. Furthermore, we also see that including interactions makes neurological and general surgery and age insignificant. We can go ahead and confirm that the non-interaction model (effect modification between type of surgery and type of admission) is not better than the non-interaction model.

Anova to show model without interaction is better than model with interaction between type of surgery and type of admission:

H_0 : All the $\beta_{interactions}$ are insignificant (i.e.$\beta_{interactions} = 0$).

H_A : Atleast one of the $\beta_{interactions}$ are significant (i.e.$\beta_{interactions} \neq 0$).

```
#Compare interaction model with non-interaction model
anova(cox_full2,cox_full3)
```

	loglik <dbl>	Chisq <dbl>	Df <int>	P(>|Chi|) <dbl>
1	-1247.392	NA	NA	NA
2	-1243.594	7.596756	8	0.4738167

2 rows

From the output above, we see that p-value 0.4738 > 0.05 (default value of alpha). Hence we accept the null-hypothesis and we can say that the there is evidence that the model without interaction in a better model.

Therefore, our final model is a multivariate cox regression model that includes the prognostic variables type of admission (urgent, emergency and elective), patient age and the main exposure variable type of surgery.

Final Multivariate Cox Regression with significant variable.

```
#Final model
final_cox = coxph(Surv(Stay_int, Delta) ~ factor(Surgery)  + factor(Type) + Age , data =
surgery, ties = "breslow")
summary(final_cox)
```

```
## Call:
## coxph(formula = Surv(Stay_int, Delta) ~ factor(Surgery) + factor(Type) +
##     Age, data = surgery, ties = "breslow")
##
##   n= 1393, number of events= 218
##
##                                        coef       exp(coef)        se(coef)
## factor(Surgery)Cardiac        -0.9764909176    0.3766304099    0.5856643977
## factor(Surgery)Neurological   -0.6550603543    0.5194107130    0.2740079544
## factor(Surgery)General        -0.6236737390    0.5359717958    0.2246347883
## factor(Surgery)Thorasic        0.2531329280    1.2880544840    0.2488445024
## factor(Surgery)Vascular      -14.2196387993    0.0000006676 1404.4910358378
## factor(Type)EMERGENCY          1.2726022667    3.5701309175    0.1974159623
## factor(Type)URGENT             0.8111544641    2.2505046144    0.4037899827
## Age                            0.0068831196    1.0069068627    0.0017227570
##                                    z       Pr(>|z|)
## factor(Surgery)Cardiac      -1.667         0.0955 .
## factor(Surgery)Neurological -2.391         0.0168 *
## factor(Surgery)General      -2.776         0.0055 **
## factor(Surgery)Thorasic      1.017         0.3090
## factor(Surgery)Vascular     -0.010         0.9919
## factor(Type)EMERGENCY        6.446 0.000000000115 ***
## factor(Type)URGENT           2.009         0.0446 *
## Age                          3.995 0.000064582526 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                exp(coef)   exp(-coef) lower .95 upper .95
## factor(Surgery)Cardiac      0.3766304099       2.6551    0.1195    1.1870
## factor(Surgery)Neurological 0.5194107130       1.9253    0.3036    0.8887
## factor(Surgery)General      0.5359717958       1.8658    0.3451    0.8324
## factor(Surgery)Thorasic     1.2880544840       0.7764    0.7909    2.0977
## factor(Surgery)Vascular     0.0000006676 1497996.0396    0.0000       Inf
## factor(Type)EMERGENCY       3.5701309175       0.2801    2.4246    5.2568
## factor(Type)URGENT          2.2505046144       0.4443    1.0199    4.9658
## Age                         1.0069068627       0.9931    1.0035    1.0103
##
## Concordance= 0.729  (se = 0.019 )
## Likelihood ratio test= 92.26  on 8 df,   p=<2e-16
## Wald test            = 68.06  on 8 df,   p=0.00000000001
## Score (logrank) test = 83.52  on 8 df,   p=1e-14
```

From the output of the final model, we see that neurological surgery, general surgery, emergency admission, urgent admission and patient age are all significant (p-value less than default value of alpha at 0.05. The p-value for all three overall tests (likelihood, Wald and logrank test) are significant, indicating that the model is significant.

Here we need to now check for assumptions to ensure our model meets the Cox PH assumption. Thus, we move into the assumption testing.

Checking for proportional hazard assumption.

For the proportional hazards to be met, the p-values for Age, Surgery (type of surgery), Type (type of admission) and the global test must be insignificant. Using the `cox.zph` function, we build the following table. Here we test the hypothesis that:

H_0 : Cox PH assumption IS valid in the model.

H_A : Cox PH assumption IS NOT valid in the model.

```
cox.zph.fit1 = cox.zph(final_cox)
cox.zph.fit1
```

```
##                   chisq df      p
## factor(Surgery)  8.00  5 0.1563
## factor(Type)    11.85  2 0.0027
## Age              3.41  1 0.0649
## GLOBAL          19.33  8 0.0132
```

From the summary table above, two covariates are statistically insignificant and even the global test is insignificant. But one level of the covariate, type of admission (Type) does not meet the Cox PH assumption. Since, the interaction above as well was not significant, we cannot introduce interactions, as sometimes it makes the proportional hazard (PH) assumption true.

Note: Something we did here to deal with this particular issue is we tried changing reference level for type of admission (since it is categorical). The hazards may be proportional when compared to one reference category but not the other. Hence, by switching the reference categories, we wanted to see which category might drive the PH assumption to be true.

After testing different methods to deal with for the assumption to be met, we were not able to get the covariate to meet the PH assumption. Since, the switching of reference levels did not help us solve our problem, this indicated that the hazards is not proportional for this particular covariates, i.e. different hazards at different time points for this covariate. If the covariate was continuous in nature, we could have applied something called as stratification of the model over different time points to be able to meet the PH assumption. Since, Type being a categorical variable, we could not make the stratification happen.

After careful examination and looking at general clinical studies closely related to the MIMIC-III data, we decided to drop the variable for types of admission (Type) as it is not a critical predictor of survival over time after patient has had their surgery. Additionally, the variable is not a biological or medical feature which would be critical in any prediction, rather it is an administrative variable used for the stratification of admission class to ICU.

Final model

```
final = coxph(Surv(Stay_int, Delta) ~ Age+Surgery, data =  surgery, ties="breslow")
summary(final)
```

```
## Call:
## coxph(formula = Surv(Stay_int, Delta) ~ Age + Surgery, data = surgery,
##     ties = "breslow")
##
##   n= 1393, number of events= 218
##
##                                coef       exp(coef)        se(coef)      z
## Age                    0.0057952884    1.0058121136    0.0017104333  3.388
## SurgeryCardiac        -1.2394435069    0.2895453031    0.5871362978 -2.111
## SurgeryNeurological   -0.8863422648    0.4121605732    0.2739101004 -3.236
## SurgeryGeneral        -0.9796629901    0.3754376037    0.2209492916 -4.434
## SurgeryThorasic       -0.2215231739    0.8012973528    0.2417080829 -0.916
## SurgeryVascular      -14.4255650205    0.0000005433 1435.2020128359 -0.010
##                       Pr(>|z|)
## Age                   0.000704 ***
## SurgeryCardiac        0.034772 *
## SurgeryNeurological   0.001213 **
## SurgeryGeneral      0.00000926 ***
## SurgeryThorasic       0.359410
## SurgeryVascular       0.991980
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                         exp(coef)   exp(-coef) lower .95 upper .95
## Age                 1.0058121136       0.9942   1.00245    1.0092
## SurgeryCardiac      0.2895453031       3.4537   0.09161    0.9151
## SurgeryNeurological 0.4121605732       2.4262   0.24094    0.7050
## SurgeryGeneral      0.3754376037       2.6636   0.24348    0.5789
## SurgeryThorasic     0.8012973528       1.2480   0.49894    1.2869
## SurgeryVascular     0.0000005433 1840531.6360   0.00000       Inf
##
## Concordance= 0.658  (se = 0.021 )
## Likelihood ratio test= 41.23  on 6 df,   p=0.0000003
## Wald test            = 30.04  on 6 df,   p=0.00004
## Score (logrank) test = 38.79  on 6 df,   p=0.0000008
```

Cox PH model assumption test

Scaled Schoenfeld Residuals

Let us test again the assumption of PH for our model using the previous hypothesis.

H_0 : Cox PH assumption IS valid in the model.

H_A : Cox PH assumption IS NOT valid in the model.

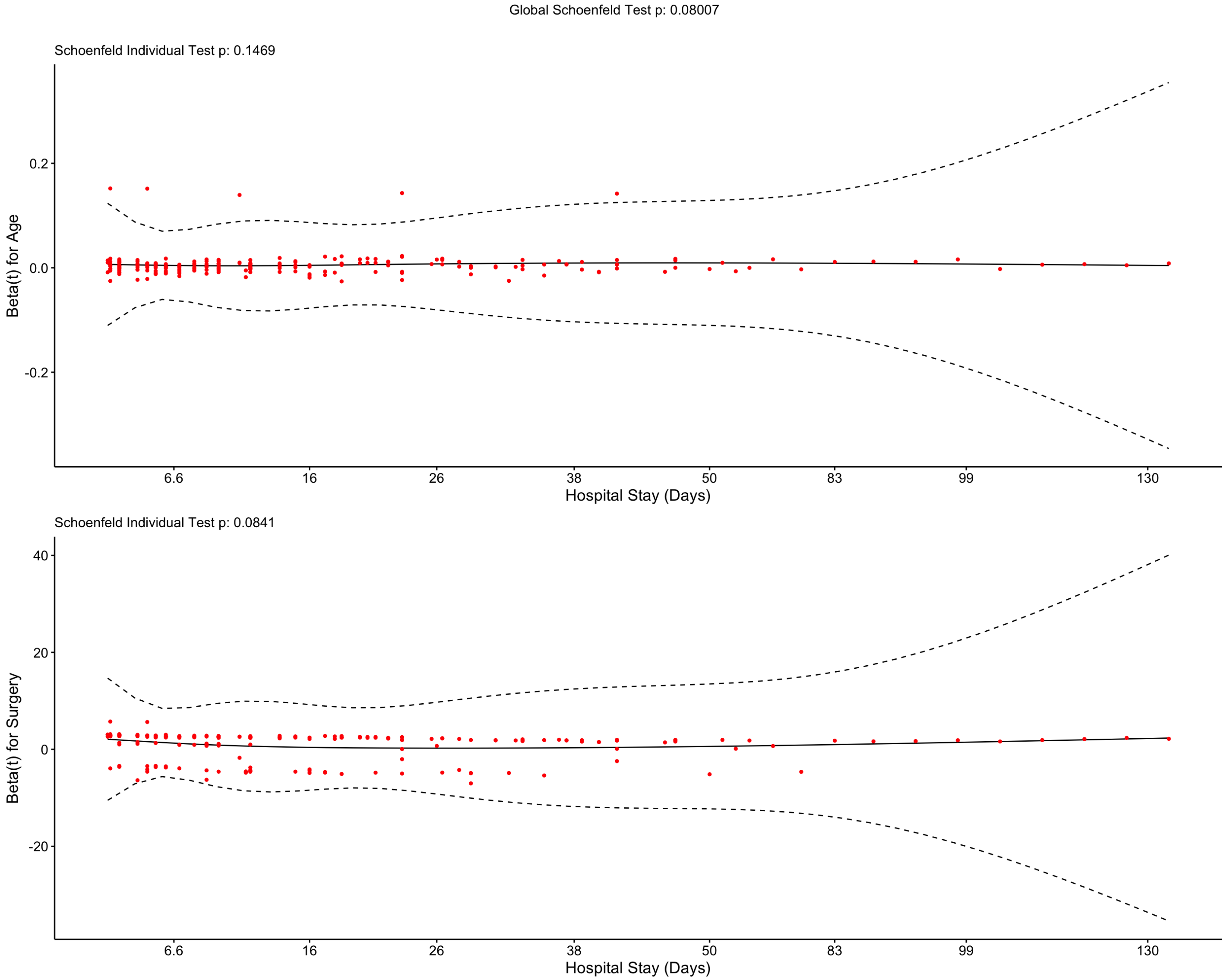
```
cox.zph.fit = cox.zph(final)
cox.zph.fit
```

```
##           chisq df     p
## Age         2.1  1 0.147
## Surgery     9.7  5 0.084
## GLOBAL     11.3  6 0.080
```

Comparing to our previous model that we had fit and tested for assumptions, this model meets all our PH assumption. Since, the global and individual test is highly insignificant, we fail to reject the null hypothesis and conclude from our hypothesis that this final model of ours which includes age as a predictor in the model with main exposure as type of surgery, meets all the PH assumptions, i.e. all the hazards are proportional for all the covariates. Hence, we can assume the proportional hazards in the model.

For testing the assumption and checking the proportional hazards graphically, we can plot a series of graphs of scaled Schoenfeld residuals for each covariate against time of observation.

```
ggcoxzph(cox.zph.fit, font.main = 12, xlab="Hospital Stay (Days)")
```



In the plot above, we have a solid smooth line and the dotted lines around the smoothing fit which represents a +/- standard-error band around that fit. From the graphical inspection, we see no patterns over time for Age and for type of surgery (Surgery). Remember, Surgery has six levels and we see a sharp deviation from the fit line for Thoracic and Vascular surgery. This means, we might have varying values for their β_s over time. If we also see the number of observation, there seems to be sparse observation of patients as we move right on the x-axis (i.e. stay longer in ICU) and that causes the confidence bands to deviate at the ends. Nonetheless, we can confirm that the assumption of proportional hazard is met.

```
summary(final)$coeff
```

```
##                              coef       exp(coef)       se(coef)           z
## Age                   0.005795288 1.0058121136100    0.001710433  3.38819903
## SurgeryCardiac       -1.239443507 0.2895453030803    0.587136298 -2.11099793
## SurgeryNeurological  -0.886342265 0.4121605731880    0.273910100 -3.23588748
## SurgeryGeneral       -0.979662990 0.3754376037134    0.220949292 -4.43388156
## SurgeryThorasic      -0.221523174 0.8012973527550    0.241708083 -0.91649055
## SurgeryVascular     -14.425565020 0.0000005433213 1435.202012836 -0.01005124
##                           Pr(>|z|)
## Age                 0.000703531937
## SurgeryCardiac      0.034772490138
## SurgeryNeurological 0.001212651751
## SurgeryGeneral      0.000009255148
## SurgeryThorasic     0.359409661568
## SurgeryVascular     0.991980402963
```

```
tab_model(final, show.reflvl = T)
```

	Surv(Stay_int, Delta)		
Predictors	*Estimates*	*CI*	*p*
Age	1.01	1.00 – 1.01	**0.001**
None	*Reference*		
Cardiac	0.29	0.09 – 0.92	**0.035**
Neurological	0.41	0.24 – 0.71	**0.001**
General	0.38	0.24 – 0.58	**<0.001**
Thorasic	0.80	0.50 – 1.29	0.359
Vascular	0.00	0.00 – Inf	0.992
Observations	1393		
R^2 Nagelkerke	0.172		

Interpretations.

Since, we checked for the assumption and established that the final model is valid, here we do the following interpretations.

A note here, a positive sign means that the hazard (risk of death) is higher, and thus the prognosis becomes worse, for subjects with higher values of that variable.

Age: Looking at the $\beta_{age} = 0.0058$ indicates that as age increases, the risk of death or hazard increases. Thus, age has a significant effect on cancer patients in our data. The hazard ratio for age i.e.,$exp(\beta_{age}) = 1.0058$ indicates that age increases the hazard of death by a factor of 1.01 as compared to a lower age. Example, a person of age 55 is 1% [`(1.01-1)*100`] more likely prone to death or has hazard ratio of 1.01 compared to someone of age 54.

Surgery: Remembering that we have 6 levels for surgery and reference level being None or no surgery. Thus, each surgery type has following effect:

- Cardiac Surgery: $\beta_{cardiac\ surgery} = -1.2394$ indicates that patients who have cardiac surgery have lower risk of death than those patients who do not have any surgery. The hazard ratio $exp(\beta_{caridac\ surgery}) = 0.2895$, means cardiac surgery has lower risk of death by a factor of 0.2895 or reduces the risk of death by 71.05% compared to patient who did not undergo surgery. We can also be 95% confident that the risk of death after cardiac surgery is atleast 0.0916 to atmost 0.915 times compared to the risk of death from no surgery at all (values from summary table).

- Neurological Surgery: $\beta_{neurological\ surgery} = -0.8863$ indicates that patients who undergo neurological surgery have lower risk of death than those patients who do not have any surgery. The hazard ratio $exp(\beta_{neurological\ surgery}) = 0.4122$, means neurological surgery reduces the risk of death by a factor of 0.4121 or 58.79%.

- General Surgery: $\beta_{general\ surgery} = -0.9797$ indicates that patients who undergo general surgery have lower risk of death than those patients who do not have any surgery. The hazard ratio $exp(\beta_{general\ surgery}) = 0.3754$, means general surgery reduces the risk of death by a factor of 0.3754 or 62.46%.

- Thorasic Surgery: $\beta_{thorasic\ surgery} = -0.2215$ indicates that patients who undergo thorasic surgery have lower risk of death than those patients who do not have any surgery. The hazard ratio $exp(\beta_{thorasic\ surgery}) = 0.8013$, means neurological surgery reduces the risk of death by a factor of 0.8012 or 19.88%.

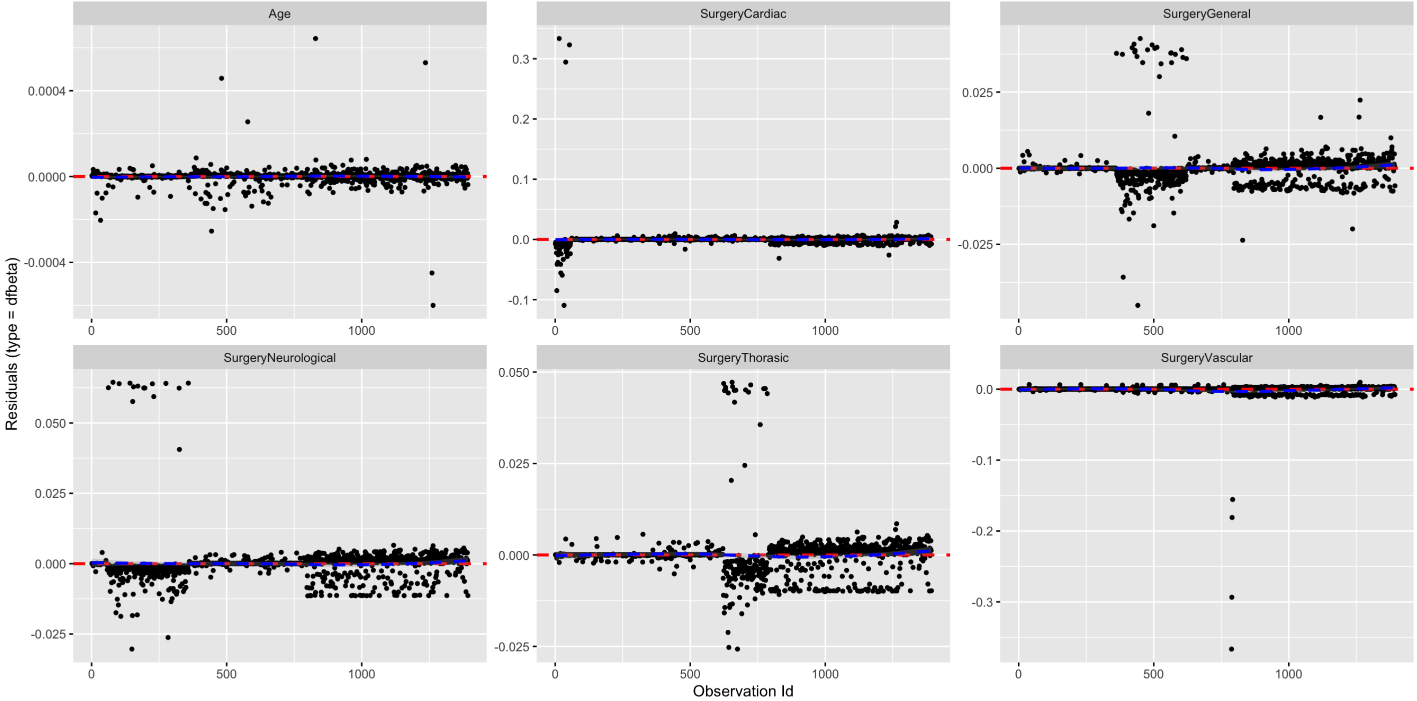
- Vascular Surgery: $\beta_{vascular\ surgery} = -14.4256$ indicates that patients who undergo vascular surgery have lower risk of death than those patients who do not have any surgery. The hazard ratio $exp(\beta_{vascular\ surgery}) = 0.000000543$, means vascular surgery reduces the risk of death by a factor of 0.000000543 or 99.9999457%.

Testing for influential subjects in the data.

In the Cox PH regression model, we can test for influential outliers or observations in our data.

```
pd = ggcoxdiagnostics(final, type = "dfbeta",
                 linear.predictions = FALSE, ggtheme = theme_grey())

plot(pd)
```



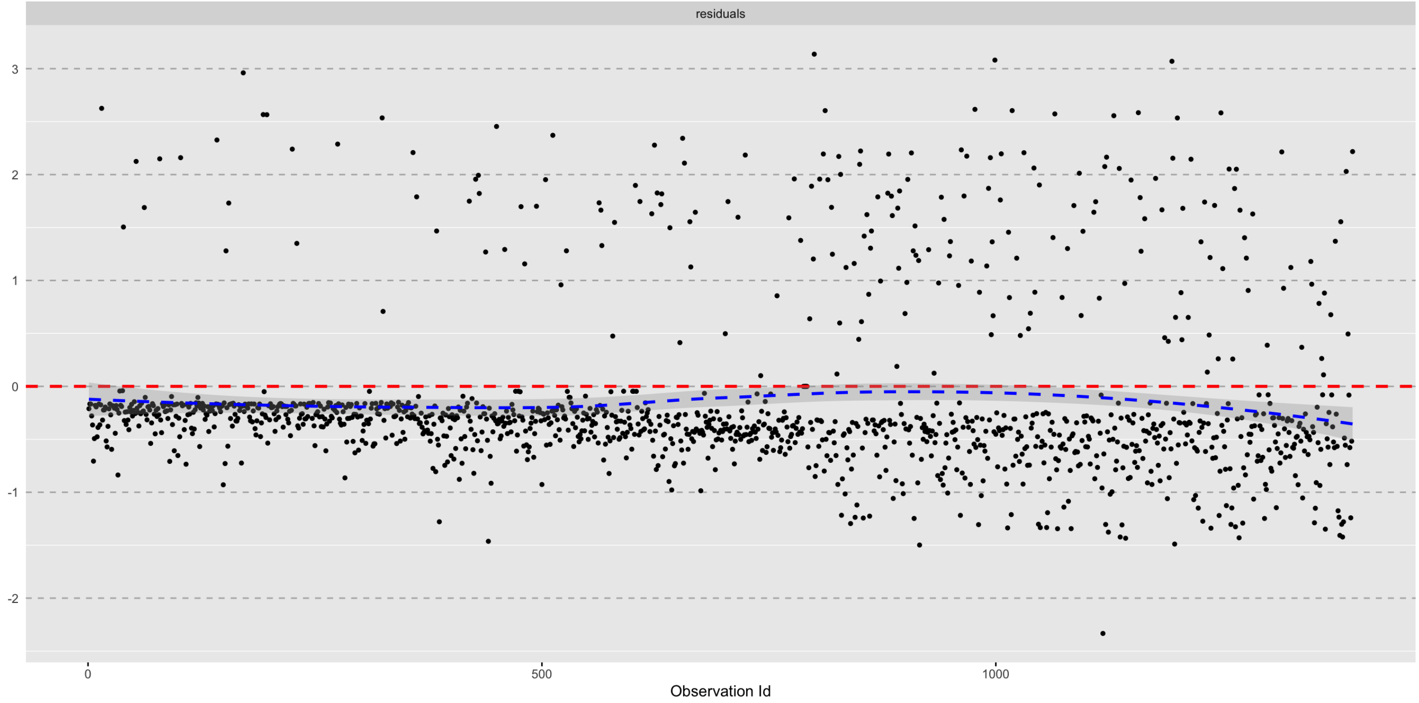
The above plots show that comparing the magnitude of the largest dfbeta values to the regression coefficients suggests that none of the patients is highly influential individually, even though some of the dfbeta values for cardiac and vascular are large compared with the other variables

Testing for the outliers using deviance residuals.

We can check for the outliers by visualizing the deviance residuals as well. The deviance residual is a normalized transformation of the martingale residual. These residuals should be roughly symmetrically distributed about zero with a standard deviation of one.

```
pr = ggcoxdiagnostics(final, type = "deviance",
                      linear.predictions = FALSE, ggtheme = theme_cleveland())

plot(pr)
```



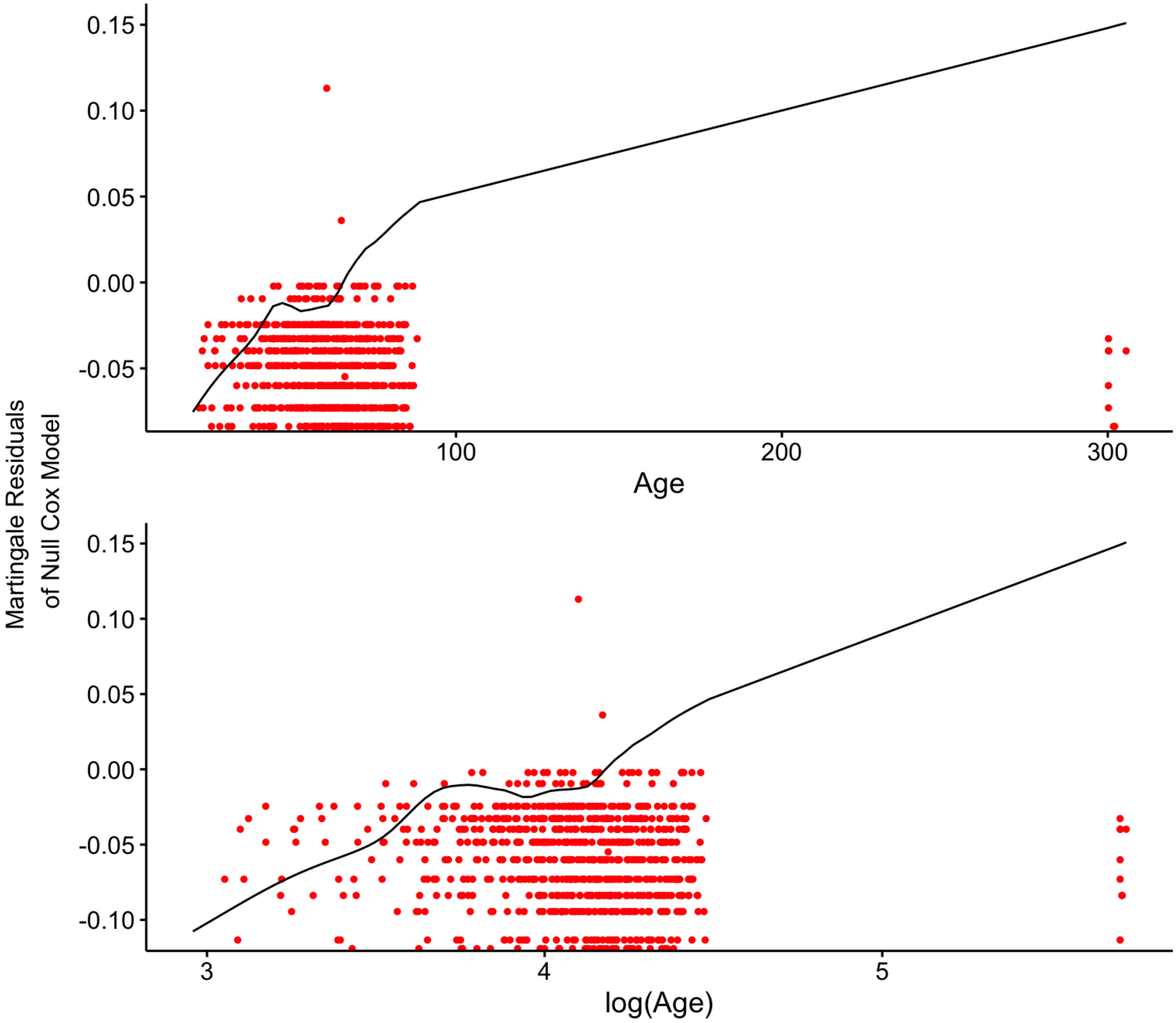
From this plot we check for deviance for the residuals of our observations. Most residuals are spread symmetrically around the zero fit line. The values over zero are related to patients who died before the expected survival time. Values below zero are related to patients who lived longer than expected survival time.

Testing for non-linearity for Age.

Often, we assume that continuous covariates have a linear form. However, this assumption should be checked. Plotting the Martingale residuals against continuous covariates is a common approach used to detect non-linearity. For a given continuous covariate, patterns in the plot may suggest that the variable is not properly fit. Non-linearity is not an issue for categorical variables, so we only examine plots of martingale residuals and partial residuals against a continuous variable.

We will check for linearity assumption for the model by the following plot.

```
ggcoxfunctional(Surv(Stay_int, Delta) ~ Age+log(Age), data =  surgery, fit = "loess")
```



From the plot above, we see that there might be some non-linearity present within our continuous covariate. We also included a log transformation for age to see its effect as well.

Results

From the detailed analysis carried above for the 1394 unique cancer patients in the dataset we created from the larger MIMIC-III dataset, we investigated and discovered the following interesting findings.

In context to predict the outcome of an event i.e. death or discharge, we found that the time spent in ICU was a significant predictor. We wanted to analyze whether time spent in ICU had an effect on the outcome and it was an important feature in the model. We also deduced that the type of admission under which the cancer patients were categorized also played a significant role, and was accompanied by age of the patient in predicting the likely outcome of survival or death.

For the survival analysis conducted on the type of admission under which the cancer patient was admitted in the ICU, we found that the patients who were admitted under the emergency category were more likely to experience the event of death than those in elective or urgent category. Also, the median hospital stay time in the emergency admission category was approximately forty days against the other categories which had median survival time of seventy days. Also, comparing the sex of the patient in the survival analysis tells us, if the patient is female, the median survival or stay in the ICU was approximately forty days, whereas for males the median length of stay was around 75 days.

Lastly, testing and analyzing the Cox Proportional Hazard Regression Model, where we were interested in seeing whether time has an effect on the outcome of death or discharge from ICU for different types of surgeries experienced by the patients. We found out that, infact surgery does improve the survival rate of the cancer patients over time after the surgery, compared to those who did not undergo any surgery. So surgery does have a significant effect on the survival of patients in our data. From the model, we also did find age to be a significant predictor for type of surgery and its effect on the overall outcome.

Discussion and Conclusions

The results above contribute to the body of literature suggesting that some form of surgery does have a benefit or good effect on the survival of cancer patient. We described the hypothesis at the beginning and conducted our analysis towards testing it. We conclude that cancer patients, who have some form of surgery and after spending some time in the ICU for recovery, do have favorable outcome i.e. have better survival rate compared to those patients who did not have a surgery. When we started off with the hypothesis, we expected to see improvements in the survival of cancer patients over an observed time period as surgeries are meant to alleviate suffering (palliative) or to remove the tumor in affected areas (preventative) of patients from a given cancer diagnosis. The conclusion of our results were very much similar to our expectation.

But in general, our expectation and results versus some other theories that have been in debate might influence the health studies differently. In general practice, the expectation of lower survival rate in patients who undergo surgery is quite prevalent in the medical field. This could be due to numerous factors and might have effects differently from different patients. This could depend on age, diagnosis, genetic build, sex, etc. There could be external factors such as socio-economic effect on the patient, health care cost, medical aid, etc. These form more of a non-biological factors for accounting in survival of a patient, in general. While studying our data and investigating our model, we realized few weakness and limitation that come from MIMIC-III dataset.

Limitations

Though our approach showed strong performance for several tasks in this dataset, this method currently has limitations in terms of generalization. These limitations might hinder or provide diluted results in further analysis and future studies. Following were some key limitations we noted while working on the dataset.

1. MIMIC-III data has the time records for different variables associated with time, censored or modified to show different time frame for potential privacy reasons. This made it challenging for us to triangulate exactly when in realistic experience were the hospital stays and events for the patients in the dataset measured or recorded.

2. Due to this reason, we were also not able to accurately factor in the time and period in which a patient had undergone a specific surgery. Due to this, we made an assumption of the surgeries to have occurred before the time spent by patients in the ICU.

3. After reading few clinical research papers, we learned that a presence of some of emergency severity index (ESI) (Louden et al.) is a common practice in medical records. The index explains the severity of an illness in the patient group between patients. This specific index is particularly viable and important for patients visiting ICU. MIMIC-III dataset, does not have such a feature available in it. To proxy this, we used type of admission to ICU, but it does not bare much evidence or value as an actual index might do. Also, type of admission is largely used to stratify patient group for administration purposes. In such future studies as the one done by us, having a severity index, would be a good predictor to use for adjusting the effect of surgeries in survival analysis. To study between our different cancer patients, addition of a form of severity index would enhance the model to provide better results.

4. Our model also has more than 700 different cancer diagnosis. It was not feasible to run the model with these many different diagnosis, some being recorded with errors or duplication for same diagnosis with different nomenclature system. This might make the analysis diluted and introduce errors in the results. Further along, some cancer has worse prognosis than the other. We did not account for different cancer types in our model but might be applicable for a modified analysis.

5. Cancer survivors suffer from many comorbid conditions that could be of great influence on their health and the effect on cancer itself. Conditions such as hypertension, hyperlipidemia, osteoarthritis, hypothyroidism, diabetes mellitus, etc. are of significant health implications and could have significant prevalence in cancer patients. This information is not easily interpretable due to amount of records in the MIMIC-III dataset with ICD-9 coding, and would need a lot more computation power than a simple laptop. Having these comorbidities readily available for the patients in the data, would help analyze their effect to our outcome in the hypothesis. This could be a consideration in further studies to apply more computation power (different programming language such as linux or C++) to access the information and decode it for the actual comorbidities rather than ICD-9 coding.

Inclusion and presence of some variables related to lab work and readings associated to blood pressure, diabetes, creatinine, etc. would be highly valuable and could be significant in our future modeling. Although the file related to blood work was available in the MIMIC data, the sheer volume and size of it, led to the file not being able to read with methods and tools available to us. Making the data more concise for elementary studies would be invaluable.

References

1. Newell, Christopher, Barbara Ramage, Alberto Nettel-Aguirre, Ion Robu, and Aneal Khan. “Peak Jump Power Reflects the Degree of Ambulatory Ability in Patients with Mitochondrial and Other Rare Diseases.” In JIMD Reports, Volume 33, pp. 79-86. Springer, Berlin, Heidelberg, 2016.

2. Beaulieu-Jones, Brett K., Patryk Orzechowski, and Jason H. Moore. “Mapping patient trajectories using longitudinal extraction and deep learning in the MIMIC-III Critical Care Database.” In PSB, pp. 123-132. 2018.

3. Rich, Jason T., J. Gail Neely, Randal C. Paniello, Courtney CJ Voelker, Brian Nussenbaum, and Eric W. Wang. “A practical guide to understanding Kaplan-Meier curves.” Otolaryngology—Head and Neck Surgery 143, no. 3 (2010): 331-336.

4. Chen, Herbert, Jeffrey M. Hardacre, Ali Uzar, John L. Cameron, and Michael A. Choti. “Isolated liver metastases from neuroendocrine tumors: does resection prolong survival?.” Journal of the American College of Surgeons 187, no. 1 (1998): 88-92.

5. Louden, B. Asher, Daniel J. Pearce, Wei Lang, and Steven R. Feldman. “A Simplified Psoriasis Area Severity Index (SPASI) for rating psoriasis severity in clinic patients.” Dermatology online journal 10, no. 2 (2004): 7-7.
6. Roy, Satyajeet, Shirisha Vallepu, Cristian Barrios, and Krystal Hunter. “Comparison of comorbid conditions between cancer survivors and age-matched patients without cancer.” Journal of clinical medicine research 10, no. 12 (2018): 911.

Appendix

The data wrangling and subsetting of cancer patients from MIMIC-III data was completed in Python. The Jupyter notebook is attached separately to see the methods and steps taken to construct the final data file for our use which is “Cancer_Surgery.csv”.

The code chunk for investigating general descriptive statistics for the data is attached below.

```
# # General Descriptive Statistics on the dataset procured from MIMIC for Cancer patients.
#
# #Loading dataset, splitting for males and females and computing mean, meadian stats for age of both sexes, hospital stay length, admission type distribtion, etc.
# cancer = read.csv("Cancer_Surgery.csv", header = TRUE)
# females = filter(cancer, Sex == 'F')
# males = filter(cancer, Sex == 'M')
# #nrow(cancer)
#
# #Calculating percentage of females in the dataset
# #nrow(males)
# #nrow(females)
# females = (667/1393)*100
# males = (726/1393)*100
#
# #Mean age of all cancer patients in the dataset
# age = cancer$Age
# mean(age)
# median(age)
#
# # Average in-hospital stay in days
# hospital_stay = cancer$Stay_int
# mean(hospital_stay)
#
# #Average hospital stay per type of admission (emergency admission, elective admission and urgent admission)
#
# #Emergency
# emergency = filter(cancer, Type == 'EMERGENCY')
#
# #Hospital stay per cancer patient whose admission is EMERGENCY
# emergency_stay = emergency$Stay_int
# mean(emergency_stay)
#
# #Elective
#
# elective = filter(cancer, Type == 'ELECTIVE')
#
# #Hospital stay per cancer patient whose admission is ELECTIVE
# elective_stay = elective$Stay_int
# mean(elective_stay)
#
# #Urgent
#
# urgent = filter(cancer, Type == 'URGENT')
#
# #Hospital stay per cancer patient whose admission is URGENT
# urgent_stay = urgent$Stay_int
# mean(urgent_stay)
#
# # In-hospital mortality
#
# total_mortality = cancer$Delta
# sum(total_mortality)
#
# inhosp.mort = (218/1393)*100
#
# # In-hospital cancer mortality startified by emergency admission
#
# emergency_mortality = emergency$Delta
# sum(emergency_mortality)
#
# can_emer.mort = (174/218)*100
#
# # In-hospital cancer motality stratified by elective admission
#
# elective_mortality = elective$Delta
# sum(elective_mortality)
#
# can_elec.mort = (36/218)*100
#
# # In-hospital cancer mortality startified by urgent admission
#
# urgent_mortality = urgent$Delta
# sum(urgent_mortality)
#
# can_urg.mort = (8/218)*100
```

Code chunk for the final logit model calculation.

```
#Calculation for final logit model
# exp(-3.560092409+0.015020433+0.006015989)/(1+exp(-3.560092409+0.015020433+0.006015989))
# exp(-3.560092409+0.015020433+0.006015989+1.957784856)/(1+exp(-3.560092409+0.015020433+0.006015989+1.957784856))
# exp(-3.560092409+0.015020433+0.006015989+1.659224184 )/(1+exp(-3.560092409+0.015020433+0.006015989+1.659224184))
```

Code chunk for logit model example 1 and 2.

```
#Calculation for example 1.
# exp(-3.560092409+0.015020433+0.006015989*(70)+1.957784856*(1))/(1+exp(-3.560092409+0.015020433+0.006015989*(70)+1.957784856*(1)))

#Calculation for example 2.
# exp(-3.560092409+0.015020433+0.0060159898*(55)+1.659224184*(1))/(1+exp(-3.560092409+0.015020433+0.006015989*(55)+1.659224184*(1)))
```