# Evaluating the Impact of Early Physiotherapy

**MATH11188 Statistical Research Skills: Scientific Report**
**Group 4 -** Aravindh Sankar Ravisankar, Nayem Dewan, Alexandros Stamatiou, Hrushikesh Vazurkar

## 1 Introduction - Setting the Stage

It is well established that prolonged bedrest can have harmful effects on patients, ranging from muscle mass decline [1, 2] to deep vein thrombosis and pulmonary thromboembolism [3]. **Early physiotherapy (PT)** intervention has been promoted as an approach to prevent some of these consequences whilst reducing the length of hospital stay and the risk of re-admission [4]. In a comprehensive review of **randomised controlled trials (RCTs)** examining early mobilization of patients who underwent cardiac surgery [5], the authors found that physiotherapy had a **positive impact** on the average **length of hospital stay**. They also concluded that the early timing of the intervention was more important than the type or intensity of the prescribed physiotherapy. We will assess these conclusions using the patient data provided to us by Hospital 1 (H1) and Hospital 2 (H2).

## 2 Data Dive - Insights from Hospital Records

The data from HP1 and HP2 were merged into a single dataset, setting "NA" for the COPD risk scores missing from HP1. Moreover, **mode imputation** was performed on **days_to_first_PT** and **PT_hours** based on patient age.

**Linear Regression:** Considering the length of stay (**LOS = days_to_discharge**) as the response variable and sex, age, cardio_risk_score, COPD_risk_score, PT_hours, days_to_first_PT as predictor variables resulted in 63 candidate models. A step-wise process of elimination was implemented for model selection. Based on minimizing the AIC, the most important determinants of LOS were found to be **age, PT_hours and days_to_first_PT**.

**Generalised Linear Model (GLM) with Poisson Link**: In comparison with the above baseline model, GLM was implemented with a similar stepwise process and the model was fitted with the resulting set of parameters listed in Table 1. The model scores were observed to be poor, making it less feasible to implement.

Table 1: Model summaries

**LINEAR MODEL:**
LOS $\sim -1 + \text{age} + \text{PT\_hours} + \text{days\_to\_first\_PT}$

| Covariate: | Estimate: | $p$-value: |
|---|---|---|
| age | $\beta_1 = 0.05 \pm 0.01$ | $< 0.001$ |
| PT_hours | $\beta_2 = -2.15 \pm 0.55$ | $< 0.001$ |
| days_to_first_PT | $\beta_3 = 1.55 \pm 0.12$ | $< 0.001$ |

**Relative Goodness-of-fit:**
AIC = 337.7, adj.$R^2 = 0.8412$
**Absolute Goodness-of-fit:**
RMSE = 1.984

**GENERALISED LINEAR MODEL:**
LOS $\sim \text{age} + \text{PT hours} + \text{days\_to\_first\_PT}$
(with Poisson link function)

| Covariate: | Estimate: | $p$-value: |
|---|---|---|
| intercept | $\beta_0 = 0.39 \pm 0.21$ | $< 0.1$ |
| age | $\beta_1 = 0.01 \pm 0.003$ | $< 0.001$ |
| PT hours | $\beta_2 = -0.54 \pm 0.14$ | $< 0.001$ |
| PT start | $\beta_3 = 0.35 \pm 0.03$ | $< 0.001$ |

**Relative Goodness-of-fit:**
AIC = 1003.9
**Absolute Goodness-of-fit:**
RMSE = 2.024

## 3 Impact of Early Physiotherapy on Recovery

Both models considered above reveal that the length of hospitalization increases with the number of days until PT is started. For instance, the linear model predicts that a 55-year-old patient starting a 1-hour daily PT treatment on day 0 can expect to be discharged after 0.67 days. On the other hand, if PT is delayed until day 3, the patient is discharged after 5.32 days. The proportionality between LOS and days_to_first_PT suggests that it is advantageous to start PT early, confirming the results from several RCTs on early mobilization (see [5] and references therein).

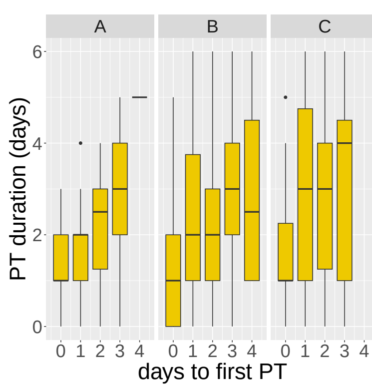


Figure 1: days_to_first_PT vs. PT duration

The selected models also predict an increase in LOS with the patient's age. To investigate this and to further evaluate the impact of early PT, we calculated the number of days of PT treatment for each patient using **PT duration = LOS − days_to_first_PT**. The raw data are represented with boxplots in Figure 1, stratified into three patient age groups: **A (< 50 years), B (50-65 years),** and **C (> 65 years)**. Within each age group, we observe that a **shorter PT treatment** can be expected if PT is started earlier. This additional benefit of early PT could be related to the idea that prolonged bedrest accelerates physical decline and then necessitates a longer period of recovery through PT treatment.

# 4 Impact of Physiotherapy Treatment Intensity

The modelling results from Section 2 show a decrease in LOS if PT hours are increased. To illustrate the impact of the PT treatment frequency on the recovery time, we considered three intensity levels: **L ('low intensity': 15-20 mins), M ('medium intensity': 25-40 mins), and H ('high intensity': 45-60 mins)**. Figure 2 shows plots of PT duration against PT intensity level for age groups A-C. The data reveal that PT duration is reduced when a high-intensity treatment is prescribed. This contradicts the findings from the RCTs considered by Santos *et al.* [5], which found that the type or frequency of mobilization had little to no impact on the length of stay. On the other hand, in a recent study of acutely hospitalized older adults, Gallardo-Gomez *et al.* [6] conclude that "optimal improvements in function are provided by either $\sim 50$ min/d of slow walking or $\sim 40$ min/d spent in multi-component interventions". This seems to echo the trends visible in the present dataset, particularly concerning age group C.

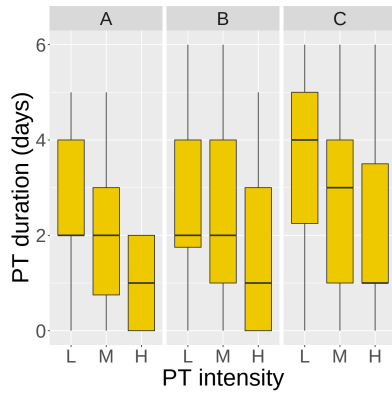

Figure 2: PT intensity vs. duration

# 5 Limitations

The datasets from HP1 and HP2 **lack** key outcome variables such as **stay in the Intensive Care Unit (ICU)**, **time of extubation**, and **incidence of post-operative complications**, all of which should be taken into account when evaluating the success of early PT [5, 7]. For instance, in a study of 7,457 patients [8] it was shown that 6.9% developed at least one postoperative complication which increased the length of hospitalization of these patients by 114% on average. It was also shown that early PT can lead to a lower incidence of complications [9].

It is **not known** which **type of cardiac surgery** was performed, which could be a further determinant of the length of hospitalization. For example, in a study by Oliveira *et al.* [10], the average time of hospital stay was $10 \pm 2.7$ days compared to $14.7 \pm 10.97$ days for patients who underwent CABG and non-CABG surgery respectively.

Another **missing** outcome variable that should be considered when evaluating early mobilization is a **patient's functional capacity** as measured through indices such as the 6-minute walking test (6MWT). In their meta-analysis of RCTs studying early mobilization after CABG surgery, Kanejima *et al.* [11] found that patients who received early PT walked 54m (95% CI 31.1m-76.9m) further at discharge than patients who did not.

# 6 Conclusion

Based on the HP1 and HP2 data, we would recommend the adoption of an early physiotherapy protocol. The formulated models reveal that starting post-operative physiotherapy earlier is expected to lead to an earlier discharge. Additionally, exploratory data analysis indicated that early physiotherapy limits the negative effects of bedrest and consequently leads to a shorter treatment duration. We also found evidence that positive results could be achieved by the prescription of a more intense physiotherapy program, echoing recent findings in the literature. Taken together, we believe that hospitals will benefit from a standardized early mobilization protocol in terms of reduced operating costs and enhanced capacity and patient recovery. In this regard, the Enhanced Recovery After Surgery (ERAS) protocol [12] should be considered as a benchmark example.

# References


1. Bloomfield, S. A. Changes in musculoskeletal structure and function with prolonged bed rest. *Medicine & Science in Sports & Exercise* **29(2),** p. 197–206 (1997).
2. English, K. L. & Paddon-Jones, D. Protecting muscle mass and function in older adults during bed rest. *Current Opinion in Clinical Nutrition and Metabolic Care* **13,** p. 34–39 (2010).
3. Dock, W. The evil sequelae of complete bedrest. *JAMA* **125(16),** p. 1083–1085 (1944).
4. Freeman, R. & Maley, K. Mobilization of intensive care cardiac surgery patients on mechanical circulatory support. *Crit Care Nurs Q* **36,** p. 73–88 (2013).
5. Santos, P. M. R., Ricci, N. A., Sustera, E. A. B., Paisani, D. M. & Chiavegato, L. D. Effects of early mobilisation in patients after cardiac surgery: a systematic review. *Physiotherapy* **103,** p. 1–12 (2017).
6. Gallardo-Gómez, D. *et al.* Optimal dose and type of physical activity to improve functional capacity and minimise adverse events in acutely hospitalised older adults: a systematic review with dose-response network meta-analysis of randomised controlled trials. *Br J Sports Med* **57(19),** p. 1272–1278 (2023).
7. Clinia, E. & Ambrosino, N. Early physiotherapy in the respiratory intensive care unit. *Respiratory Medicine* **99,** p. 1096–1104 (2005).
8. Khan, N. A. *et al.* Association of postoperative complications with hospital costs and length of stay in a tertiary care center. *J Gen Intern Med* **21,** p.177–80 (2006).
9. Herdy, A. H. *et al.* Pre- and postoperative cardiopulmonary rehabilitation in hospitalized patients undergoing coronary artery bypass surgery. *Am J Phys Med Rehabil* **87,** p. 714–9 (2008).
10. Oliveira, G. U. *et al.* Determinants of distance walked during the six-minute walk test in patients undergoing cardiac surgery at hospital discharge. *Journal of Cardiothoracic Surgery* **9(95),** p. 1146–1161 (2014).
11. Kanejima, Y., Shimogai, T., Kitamura, M., Ishihara, K. & Izawa, K. P. Effect of Early Mobilization on Physical Function in Patients after Cardiac Surgery: A Systematic Review and Meta-Analysis. *Int. J. Environ. Res. Public Health 2020, 17, 7091* **17,** p. 7091 (2020).
12. Petersen, J. *et al.* Economic impact of enhanced recovery after surgery protocol in minimally invasive cardiac surgery. en. *BMC Health Serv Res.* doi: (Mar. 20, 2021).

# A Appendix

## A.1 R Markdown File

```
---
title: "SRS 3 Debugging"
author: "Hrushikesh Vazurkar"
date: "`r Sys.Date()`"
output:
  latex_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
rm(list = ls(all = TRUE))
```

```{r}
library(dplyr)
library(tidyr)
library(ggplot2)
library(GGally)
library(tidyverse)
library(autoReg)
library(leaps)
library(ggpubr)
```

**1. Data Preprocessing**

Load data for both hospitals.

```{r}
hospital_1 <- read.csv("hospital_1_data.csv")
hospital_2 <- read.csv("hospital_2_data.csv")
```

Combine data from both hospitals into single dataframe. However, added an extra column
    hospital_id for differentiation of source hospital.

```{r}
#Add a hospital ID before merging if you need to retain the source information
hospital_1 <- hospital_1 %>% mutate(hospital_id = 1)
hospital_2 <- hospital_2 %>% mutate(hospital_id = 2)

#Assuming hospital_2 has an extra column 'COPD risk score', which hospital_1 does not
    have
hospital_1$COPD_risk_score <- NA #Add the missing column to hospital_1 with NA values

#Combine the datasets
combined_data <- bind_rows(hospital_1, hospital_2)
```

### Handle missing values - days_to_first_PT and PT_hours (Mode Imputation)

```{r}
na_counts <- colSums(is.na(combined_data))
print("Number of NA values in each column:")
print(na_counts)
```

#### Impute days_to_first_PT NA values based on patient's age

1. Find max and min ages of patients.

```{r}

min_age <- min(combined_data['age'])
max_age <- max(combined_data['age'])

min_age
max_age
```
```

````
2. Create age buckets of size 10 covering min and max ages.

```{r}
breaks <- seq(min_age,max_age+10, by = 10)
breaks
```
3. Assign age buckets for each row.

```{r}

labels <- paste0("[", breaks[-length(breaks)] + 1, "-", breaks[-1], "]")

combined_data$age_bucket <- cut(combined_data$age, breaks = breaks, labels = labels, include.lowest = TRUE)

print(combined_data[c("age","age_bucket")])
```

Age Bucket Classification :

[39 - 48] ----- 1
[49 - 58] ----- 2
[59 - 68] ----- 3
[69 - 78] ----- 4
[79 - 88] ----- 5

4. Get indices where days_to_first_PT is NA in combined_data.

```{r}
na_indices_days_to_first_PT <- which(is.na(combined_data$days_to_first_PT))
na_indices_days_to_first_PT
```
5. Get the days_to_first_PT frequency tables for each age bucket.

```{r}

age_bucket_splits <- split(combined_data,combined_data$age_bucket)

# Iterate over each age_bucket value
for (age_bucket_value in names(age_bucket_splits)) {
    cat("Age Bucket:", age_bucket_value, "\n")

    # Get unique values of 'days_to_first_PT' for the current age_bucket value
    unique_days <- table(age_bucket_splits[[age_bucket_value]]$days_to_first_PT)

    # Print the unique values and their corresponding counts
    print(unique_days)
    cat("\n")
}

```
```{r}
combined_data[57,]
```

Example - For a patient on index 57, we get the age of 76, which corresponds to age_bucket [70-79]. Post that, we use the unique_days frequency table to assign value of 1, as it is the mode in [70-79].

Same procedure is repeated for all remaining NA values in days_to_first_PT.

```{r}

combined_data[57, "days_to_first_PT"] <- as.numeric(1)
combined_data[79, "days_to_first_PT"] <- as.numeric(2)
combined_data[159, "days_to_first_PT"] <- as.numeric(1)
combined_data[205, "days_to_first_PT"] <- as.numeric(1)
combined_data[232, "days_to_first_PT"] <- as.numeric(2)
```

#### Impute Pt_hours NA values based on patient's age

````

````
The procedure is similar as imputing days_to_first_PT. However, PT_hours is a much more significant chunk as it includes 32 records (12% of the total records).

```{r}

age_bucket_splits <- split(combined_data,combined_data$age_bucket)

# Iterate over each age_bucket value
for (age_bucket_value in names(age_bucket_splits)) {
    cat("Age Bucket:", age_bucket_value, "\n")

    # Get unique values of 'days_to_first_PT' for the current age_bucket value
    unique_days <- table(age_bucket_splits[[age_bucket_value]]$PT_hours)

    # Print the unique values and their corresponding counts
    print(unique_days)
    cat("\n")
}

```

```{r}
na_indices_PT_hrs <- which(is.na(combined_data$PT_hours))
na_indices_PT_hrs
```

```{r}

for(i in na_indices_PT_hrs){
  age <- combined_data[i, "age"]
  if(age>=40 & age<=49)
    combined_data[i, "PT_hours"] <- 0.5
  if(age>=50 & age<=59)
        combined_data[i, "PT_hours"] <- 0.5

  if(age>=60 & age<=69)
        combined_data[i, "PT_hours"] <- 0.5

  if(age>=70 & age<=79)
        combined_data[i, "PT_hours"] <- 0.5

  if(age>=80 & age<=89)
        combined_data[i, "PT_hours"] <- 0.666666666666667

}

```

After these steps, the only NA values remaining are COPD_risk_score, which is by design due to merging of Hospitals 1 and 2 datasets.

### Data Stratification

#### Based on cardio_risk_score: [<=2] as low, [3,4] as medium and 5 as high risk.

```{r}
#Add a cardio risk group variable: low, medium, high
combined_data_low <- combined_data %>% filter(cardio_risk_score<=2) %>% mutate(risk_group = "low")
combined_data_med <- combined_data %>% filter(between(cardio_risk_score,3,4)) %>% mutate(risk_group = "medium")
combined_data_high <- combined_data %>% filter(cardio_risk_score==5) %>% mutate(risk_group = "high")
#merge together
combined_data <- bind_rows(combined_data_low, combined_data_med, combined_data_high)
```

#### Based on age

```{r}
#add age groups
combined_data_A <- combined_data %>% filter(age<=50) %>% mutate(age_group = "A")
````
````

````
combined_data_B <- combined_data %>% filter(between(age,51,65)) %>% mutate(age_group = "B")
combined_data_C <- combined_data %>% filter(age>65) %>% mutate(age_group = "C")
#merge together
combined_data <- bind_rows(combined_data_A, combined_data_B, combined_data_C)
```

#### Based on PT_hours

```{r}
#add PT intensity
combined_data_level1 <- combined_data %>% filter(PT_hours<=0.34) %>% mutate(PT_intensity = "L")
combined_data_level2 <- combined_data %>% filter(between(PT_hours,0.34,0.67)) %>% mutate(PT_intensity = "M")
combined_data_level3 <- combined_data %>% filter(PT_hours>0.67) %>% mutate(PT_intensity = "H")
#merge together
combined_data <- bind_rows(combined_data_level1, combined_data_level2, combined_data_level3)
```

### Adding a derived column of PT_duration

```{r}
#Let's add an extra column called PT_duration
combined_data <- combined_data %>% mutate(combined_data, PT_duration = days_to_discharge - days_to_first_PT)
```

```{r}
#convert to factors as appropriate
combined_data$sex <- as.factor(combined_data$sex)
combined_data$risk_group <- factor(combined_data$risk_group, levels = c("low","medium","high"))
combined_data$PT_intensity <- factor(combined_data$PT_intensity, levels = c("L","M","H"))
combined_data$cardio_risk_score <- as.factor(combined_data$cardio_risk_score)
combined_data$COPD_risk_score <- as.factor(combined_data$COPD_risk_score)
combined_data$age_group <- as.factor(combined_data$age_group)

#separate HP2 dataset
hp2data <- filter(combined_data,hospital_id==2)

str(combined_data)

```

**2. EDA - Exploratory Data Analysis**

## Correlations

-   First let's check how the length of the PT treatment (`PT_duration`) is related to the variables `age`, `sex`, `cardio_risk_score` and `days_to_first_PT` :

```{r}
correlations1 <- combined_data %>%
  select(sex, age, cardio_risk_score, PT_duration, days_to_first_PT) %>%
  GGally::ggpairs(aes(color=sex), columns = c("age", "cardio_risk_score", "PT_duration", "days_to_first_PT")) +
  scale_colour_manual(values = c("magenta","skyblue")) +
  scale_fill_manual(values = c("magenta","skyblue")) + theme(plot.title = element_text(hjust = 0.5))

correlations1
```

-   There is weak positive correlation between `age` and `PT_duration` , both overall and within-sex.

-   There is negligible correlation between `age` and `days_to_first_PT` and between `age` and `cardio_risk_score`

````
````

~~~
-   The is some positive correlation between `cardio_risk_score` and `days_to_first_PT`, so the reason for starting PT later could be related to the cardio risk. This is more pronounced in women.

-   A positive correlation exists also between `days_to_first_PT` and `PT_duration`. I think this fact is quite interesting. It seems to suggest that delaying the start of the PT leads to a longer PT treatment. To examine this further, let's first also consider the correlations between `PT_duration` , `PT_hours`, `days_to_discharge`, `days_to_first_PT`

```{r}
correlations2 <- combined_data %>%
  select(sex, PT_hours, days_to_discharge, days_to_first_PT, PT_duration) %>%
  GGally::ggpairs(aes(color=sex), columns = c("PT_hours", "days_to_discharge", "days_to_first_PT", "PT_duration")) +
  scale_colour_manual(values = c("magenta","skyblue")) +
  scale_fill_manual(values = c("magenta","skyblue")) + theme(plot.title = element_text(hjust = 0.5))

correlations2
```

-   First notice the large correlation between `PT_duration` and `days_to_discharge` ! This shouldn't be a big surprise: for any given patient, a longer PT treatment will obviously lead to a later discharge date.

-   The second largest correlation is between `days_to_first_PT` and `days_to_discharge`. To simplify, let's **assume** **constant** PT treatment length. Then, a later start date later will obviously lead to a later discharge data. In that case, the natural question is: **why would the PT have to be started later for some patients?** **If there is no obvious reason, it should be advocated to start as early as possible.** But we saw above that there is a possible link between the start time of PT and the cardio risk score.... So the issue isn't as simple, because it could be that there is a real physical reason for why the PT cannot start earlier.

-   What we could look at as well is **the effect of starting PT early on the duration of the treatment**. If we can show from the data that *early PT* leads to a *shorter treatment* and therefore to a shorter hospital stay, then I believe it to be an additional argument for its implementation. This can then be backed up with literature. Furthermore, we can investigate the effect of the PT intensity (`PT_hours`) on the duration. Looking at the correlations, it seems that increasing `PT_hours` could lead to a reduction of `PT_duration`.

-   But we must decouple the question from the risk score. In what follows, I will divide the patients according to their cardio risk. Within each risk group, I will then plot `days_to_first_PT` vs. `PT_duration` as well as `PT_hours` vs. `PT_duration`

### Investigating within age groups and cardio risk groups:

Let's first divide the patients into different the cardio risk groups, as the risk score could be a factor that delays the start of physiotherapy. I.e. we would expect people with higher risk score to start physiotherapy later, because of the possibility of complications. Let's examine it with a plot of risk score vs. start day of PT:

```{r}
plot(hp2data$COPD_risk_score,hp2data$risk_group)
```

```{r}
cardio_risk_v_PTstart <- ggplot(combined_data, aes(x = risk_group, y = days_to_first_PT)) + geom_boxplot(aes(fill = risk_group))

#also consider COPD risk score
COPD_risk_v_PTstart <- ggplot(hp2data, aes(x = COPD_risk_score, y = days_to_first_PT)) + geom_boxplot(aes(fill= COPD_risk_score))

cardio_risk_v_PTstart; COPD_risk_v_PTstart
```

-   As we can see, there may be a delay in the start of PT for the higher cardio risk categories (4,5) compared to the lower risk categories (1,2,3), but COPD risk score
~~~

````
does not seem to delay PT start. So cardio risk score could be a factor in the start date of PT. To analyse the effect of start date on discharge date, it therefore makes sense to **divide the patients into the different risk groups**.

-   Within each risk category, we can then examine the influence of `PT_hours` and `days_to_first_PT` on the length of the hospital stay, measured either through `days_to_discharge` or `PT_duration`.

```{r}
#five cardio risk groups
cardio1_group <- filter(combined_data,cardio_risk_score==1)
cardio2_group <- filter(combined_data,cardio_risk_score==2)
cardio3_group <- filter(combined_data,cardio_risk_score==3)
cardio4_group <- filter(combined_data,cardio_risk_score==4)
cardio5_group <- filter(combined_data,cardio_risk_score==5)

#three cardio risk groups
cardio_low_risk <- bind_rows(cardio1_group,cardio2_group)
cardio_medium_risk <- bind_rows(cardio3_group,cardio4_group)
cardio_high_risk <- bind_rows(cardio5_group)
```

```{r}
#sample sizes
#nrow(cardio_low_risk);nrow(cardio_medium_risk);nrow(cardio_high_risk)

#average days to first PT
mean(cardio_low_risk$days_to_first_PT, na.rm = T)
mean(cardio_medium_risk$days_to_first_PT, na.rm=T)
mean(cardio_high_risk$days_to_first_PT, na.rm=T)

#average days to discharge
mean(cardio_low_risk$days_to_discharge, na.rm = T)
mean(cardio_medium_risk$days_to_discharge, na.rm=T)
mean(cardio_high_risk$days_to_discharge, na.rm=T)
```

#### `days_to_first_PT` vs. `days_to_discharge`:

```{r}
#low riskgroup
PTstart_v_discharge_low_cardio <- ggplot(cardio_low_risk, aes(x = days_to_first_PT, y = days_to_discharge)) + geom_boxplot(aes(fill = as.factor(days_to_first_PT))) + geom_smooth(method="lm")

#medium riskgroup
PTstart_v_discharge_medium_cardio <- ggplot(cardio_medium_risk, aes(x = days_to_first_PT, y = days_to_discharge)) + geom_boxplot(aes(fill = as.factor(days_to_first_PT))) + geom_smooth(method="lm")

#very high riskgroup 5
PTstart_v_discharge_high_cardio <- ggplot(cardio_high_risk, aes(x = days_to_first_PT, y = days_to_discharge)) + geom_boxplot(aes(fill = as.factor(days_to_first_PT))) + geom_smooth(method="lm")

PTstart_v_discharge_low_cardio; PTstart_v_discharge_medium_cardio;PTstart_v_discharge_high_cardio
```

```{r}
#stratified by age group
ggplot(combined_data, aes(x = days_to_first_PT, y = days_to_discharge)) + geom_boxplot(aes(fill=as.factor(days_to_first_PT))) + scale_fill_manual(values = rep("red",6)) + geom_smooth(method="lm") + facet_wrap(~age_group) + theme(legend.position = "none") + labs(x="PT start day", y="Days to discharge")

#stratified by risk group
ggplot(combined_data, aes(x = days_to_first_PT, y = days_to_discharge)) + geom_boxplot(aes(fill=as.factor(days_to_first_PT))) + scale_fill_manual(values = rep("red",6)) + geom_smooth(method="lm") + facet_wrap(~risk_group) + theme(legend.position = "none") + labs(x="PT start day", y="Days to discharge")
```

````

```
-   Within each risk group, there seems to be a linear relationship between `days_to_first_PT` and `days_to_discharge`. If it were indeed linear, then the PT treatment duration is constant. This would suggests that starting the PT earlier **does not have a negative impact** on the length of the treatment and therefore should be adopted. So, unless there is a very specific reason for delaying, this should not be done. I.e. the extra few days of bedrest do not lead to a shorter treatment.

#### `days_to_first_PT` vs. `PT_duration`

```{r}
#stratified by riskgroup
ggplot(combined_data, aes(x = days_to_first_PT, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(days_to_first_PT))) + scale_fill_manual(values = rep("red",6)) + geom_smooth(method="lm") + facet_wrap(~risk_group) + theme(legend.position = "none") + labs(x="PT start day", y="PT duration")

#stratified by agegroup
ggplot(combined_data, aes(x = days_to_first_PT, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(days_to_first_PT))) + scale_fill_manual(values = rep("seagreen3",6)) + geom_smooth(method="lm") + facet_wrap(~age_group) + theme(legend.position = "none") + labs(x="PT start day", y="PT duration")

#overall
ggplot(combined_data, aes(x = days_to_first_PT, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(days_to_first_PT))) + scale_fill_manual(values = rep("red",6)) + theme(legend.position = "none") + labs(x="PT start day", y="PT duration")
```

-   I think this result is quite revealing: in the low/medium cardio risk group (1,2,3), the starting time of PT has little impact on the duration. In the higher groups, a *later* start of PT leads to a *longer* treatment. This is particularly true in the highest risk group (5). Here, starting on day 3 corresponds to an average of 4 days of physiotherapy, whereas starting on day 2, it's only 2 days of PT! This has two clear benefits: **(a) shorter PT treatment** and **(b) earlier PT start**. Both contribute to a shorter hosptital stay!

-   Can we explain it with medical reasons? I think so: longer bedrest has an adverse effect on the body (this fact is well established in the literature). It can further be presumed that the physical condition of high risk (4,5) patients is a *worse* than that of low risk patients (1,2,3). Therefore, it seems logical that the effects of prolonged bedrest are *more severe* for the high risk patients, and that a *longer therapy* is necessary to recover. I.e. the "normal" amount of PT days doesn't suffice if you wait too long! On the other hand for the lower risk groups, the deterioration due to a few days of bedrest is not as severe and this would explain why a "standard" amount of PT is still sufficient. Overall, I think these plots shows that to an early start of PT has can have positive effect on the recovery time, but **only** in the higher risk patients.

#### `PT_intensity` vs. `days_to_discharge`:

```{r}
#stratified by age group
ggplot(combined_data, aes(x = PT_intensity, y = days_to_discharge)) + geom_boxplot(aes(fill=as.factor(PT_intensity))) + scale_fill_manual(values = rep("red",6)) + geom_smooth(method="lm") + facet_wrap(~age_group) + theme(legend.position = "none") + labs(x="PT intensity", y="Days to discharge")

#stratified by risk group
ggplot(combined_data, aes(x = PT_intensity, y = days_to_discharge)) + geom_boxplot(aes(fill=as.factor(PT_intensity))) + scale_fill_manual(values = rep("red",6)) + geom_smooth(method="lm") + facet_wrap(~risk_group) + theme(legend.position = "none") + labs(x="PT intensity", y="Days to discharge")
```

#### `PT_intensity` vs. `PT_duration`:

```{r}
#stratified by age group
ggplot(combined_data, aes(x = PT_intensity, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(PT_intensity))) + scale_fill_manual(values = rep("red",6)) + geom_smooth(method="lm") + facet_wrap(~age_group) + theme(legend.position = "none") + labs(x="PT intensity", y="PT duration")

```
```

````
#stratified by risk group
ggplot(combined_data, aes(x = PT_intensity, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(PT_intensity))) + scale_fill_manual(values = rep("red",6)) + geom_smooth(method="lm") + facet_wrap(~risk_group) + theme(legend.position = "none") + labs(x="PT intensity", y="PT duration")

#overall
ggplot(combined_data, aes(x = PT_intensity, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(PT_intensity))) + scale_fill_manual(values = rep("red",6)) + geom_smooth(method="lm") + theme(legend.position = "none") + labs(x="PT intensity", y="PT duration")
```

-   In these plots we can see that the intensity of PT treatment does seem to affect the duration of the PT treatment somewhat: high-intensity treatment results in fewer treatment days compared to low-intensity treatment

-   Medically, this is supported in the literature through randomized control trials that did not notice any difference in the type of PT (here we have different intensities as "types"). So again, the results seem to confirm what is in the Santos paper: **what matters most is to start PT early, not the type/intensity of PT.** This is good for the hospitals in terms of keeping down treatment costs.

-   However: is there a slight downward trend in the means of `PT_duration` with increasing `PT_hours` , both in the low and high risk groups? If that is true, then we should advocate an increase in the hours of PT together with an early start of PT.

#### Final plots for report

```{r}
#PLOT 1
#stratified by age group
#pdf("PT_intensity_v_duration.pdf")
ggplot(combined_data, aes(x = PT_intensity, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(PT_intensity))) + scale_fill_manual(values = rep("gold2",6)) + geom_smooth(method = "lm", se = TRUE) + facet_wrap(~age_group) + theme(legend.position = "none") + labs(x="PT intensity", y="PT duration") + ylim(0,6) + theme(plot.title = element_text(hjust=0.5), legend.position = "none", text = element_text(size = 30))
#dev.off()


#PLOT 2
#pdf("PT_start_v_duration.pdf")
#stratified by agegroup
ggplot(combined_data, aes(x = days_to_first_PT, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(days_to_first_PT))) + scale_fill_manual(values = rep("gold2",6)) + facet_wrap(~age_group) + theme(plot.title = element_text(hjust=0.5), legend.position = "none", text = element_text(size = 30)) + labs(x="days to first PT", y="PT duration (days)") + ylim(0,6) + xlim(-0.5,4.5)
#dev.off()
```

**3. Modelling**

### Linear Models

#### Exhaustive feature selection through regsubsets

```{r}
library(leaps)
# Using regsubsets to explore all possible models without intercept
subset_selection <- regsubsets(days_to_discharge ~ sex + age + PT_hours + cardio_risk_score + COPD_risk_score + days_to_first_PT, intercept = TRUE, data=combined_data, nbest=1, really.big=TRUE)

# View the best models of each size based on an information criterion like BIC or AIC
canditate_models <- summary(subset_selection)
canditate_models
```

```{r}
# Inspect adjusted R^2 and BIC
````

~~~
canditate_models$adjr2
canditate_models$aic
canditate_models
```

#### Test Models

```{r}
lm_ <- lm(days_to_discharge ~ -1 + sex + age + PT_hours + cardio_risk_score + days_to_first_PT, combined_data)
step(lm_)
```

```{r}
summary(lm_)
```

```{r}
lm0 <- lm(days_to_discharge ~ sex + age + PT_hours + cardio_risk_score + days_to_first_PT, combined_data)
step(lm0)
```

#### Optimal Model as per the feature selection (without intercept)

```{r}
lmfit <- lm(days_to_discharge ~ -1 + age + PT_hours + days_to_first_PT, combined_data)
summary(lmfit)
```

```{r}
LOS_lm <- function(age, PT_hours, days_to_first_PT){
  #coefficients
  b1 <- lmfit$coefficients[1]
  b2 <- lmfit$coefficients[2]
  b3 <- lmfit$coefficients[3]
  #return prediction
  b1*age + b2*PT_hours + b3*days_to_first_PT
}


LOS_lm(55,1,3)-LOS_lm(55,1,2)
LOS_lm(55,1,2)-LOS_lm(55,1,1)
```

The below code generates image pdfs.

```{r}

#stratified by age group
pdf("PT_intensity_v_duration.pdf")
ggplot(combined_data, aes(x = PT_intensity, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(PT_intensity))) + scale_fill_manual(values = rep("gold2",6))  + facet_wrap(~age_group) + theme(legend.position = "none") + labs(x="PT intensity", y="PT duration") + ylim(0,6) + theme(plot.title = element_text(hjust=0.5), legend.position = "none", text = element_text(size = 30))
dev.off()


pdf("PT_start_v_duration.pdf")
#stratified by agegroup
ggplot(combined_data, aes(x = days_to_first_PT, y = PT_duration)) + geom_boxplot(aes(fill=as.factor(days_to_first_PT))) + scale_fill_manual(values = rep("gold2",6)) + facet_wrap(~age_group) + theme(plot.title = element_text(hjust=0.5), legend.position = "none", text = element_text(size = 20)) + labs(x="PT start (days)", y="PT duration (days)") + ylim(0,6)
dev.off()
```

### Generalised Linear Models

#### Test Model

~~~

```
```{r}
glm1 <- glm(days_to_discharge ~ age + PT_hours  + days_to_first_PT + sex + cardio_risk_score, family = poisson, data = combined_data)
summary(glm1)
step(glm1)
```

#### Optimal Model

```{r}
glm_final <- glm(days_to_discharge ~ age + PT_hours  + days_to_first_PT, family = poisson, data = combined_data)
summary(glm_final)
```

**4. Absolute Goodness-of-fit**

### RMSE for Linear Model (lmfit)

```{r}
# For linear regression model
library(Metrics)
predicted_values_linear <- predict(lmfit)
rmse_linear <- rmse(combined_data$days_to_discharge, predict(lmfit))
rmse_linear
```

### RMSE for Generalised Linear Model (glm_final)

```{r}

# For generalized linear model (e.g., Poisson regression)
predicted_values_glm <- predict(glm_final, type = "response")
residuals_glm <- combined_data$days_to_discharge - predicted_values_glm
rmse_glm <- sqrt(mean(residuals_glm^2))
rmse_glm
```
```

Listing 1: R Markdown File