Clustering

Team 1

5/5/2020

DATA CLEANING

Our data cleaning process is basically divided into 3 steps. Firstly, we looked into each column, detected and dealt with unusual values. Secondly, we determined the effectiveness of each therapeutic process based on the evaluation criteria of each outcome type. Specifically, we evaluated the effectiveness of pain reduction and the overall improvement. Then, we aggregated some therapy summaries which were logically duplicated, and reevaluated them by taking average.

I

In the first step, we mainly worked on fixing typos, removing repeated rows, detected and fixed (or removed) problematic values. Typos are general problems caused by inconsistent upper/lower case spelling. And there are many special problems for each column, we analyzed and determined how to smooth the potential errors:

(a) We deleted the record with discharge date in June 2020.

(b) We deleted a 7-digit ROMS ID.

(c) We deleted the negative age and adjusted the inconsistent age based on earlier therapy record.

(d) We adjusted chronic pain as NA if values for chronic pain conflicted with itself.

(e) We fixed inconsistent admission date and discharge date. Particularly, holding other criteria constant (ID, classification, body region, outcome and admission date), if the discharge date varied, then we adjusted all the discharge dates to be the same as the last discharge date of this therapy. Similarly, holding other criteria constant (ID, classification, body region, outcome and discharge date), if the admission date varied, then we adjusted all the admission dates to be the same as the first admission date of this therapy.

(f) After the data cleaning steps mentioned above, we grouped the data by conditions including ROMS ID, Body Region, Outcome, Classification, Admission Date and Discharge Date. We defined each group of data as a combination of therapy records during one entire treatment process. Then we recounted visit times and the time(days) during one entire therapy.

(g) We removed the outcome scores that were out of range. The maximum possible value for LOWER EXTREMITY FUNC SCALE is 80, we eliminated the records for those LOWER EXTREMITY FUNC SCALE values larger than 80 (applied to either admission scores or discharge scores).

II

In the second step, we evaluated the effectiveness of pain reduction and the overall improvement of each treatment and gave binary scores to them.

(a) For pain result, if pain scores were reduced by at least 2 points, or reduced from less than 2 to 0, then pain reduction effectiveness was good, we assigned 1 to the result. Otherwise, we considered the therapy's effectiveness of pain reduction was not good, and we assigned 0 to the result.

(b) For overall outcome result, we needed to specify different criteria for different outcome types:

   (1) When outcome type was LOWER EXTREMITY FUNC SCALE, if the score increased at least 9 points or increased from larger than 71 to 80 (max), then we considered the treatment is effective

overall, and we assigned the overall effect with score 1. Otherwise, we gave the overall effect with score 0.

(2) when outcome type was KNEE OUTCOME SURVEY, if the score increased at least 9 points or increased from larger than 91 to 100 (max), then we considered the treatment is effective overall, and we assigned the overall effect with score 1. Otherwise, we gave the overall effect with score 0.

(3) When outcome belongs to other 3 types, if the score decreased at least 10 points or decreased from less than 10 to 0 (minimum), then we considered the treatment is effective overall, and we assigned the overall effect with score 1. Otherwise, we gave the overall effect with score 0.

III

After those 2 steps, then we aggregated some therapy summaries and evaluated them by taking average. For the treatments with identical same ID, body region, classification and treatment time, if the categories of outcome are different, we took the average of the pain reduction effectiveness scores (1 or 0) and took the average of the effect for overall improvement (1 or 0). Since we aggregated some rows, some columns became invalid logically, so they are assigned NA.

```
dat1 <-  read_excel("~/Desktop/For Masanao Class - ROMS Full Data Set - March 19th, 2019 Upload.xlsx",sh
dat2 <-  read_excel("~/Desktop/For Masanao Class - ROMS Full Data Set - March 19th, 2019 Upload.xlsx",sh
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i =
## sheet, : Expecting numeric in Z10703 / R10703C26: got 'NULL'
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i =
## sheet, : Expecting numeric in Z16107 / R16107C26: got 'unknown'
```

```
# delete the weird ID
dat2<- dat2 %>% filter(`ROMS ID` != 1000330)


# fix inconsistent admission date and discharge date.
# (it may merge some patients with relapse issues.)
dat2$`Visit Date` <- as.character(dat2$`Visit Date`)
dat2$`Admission Date`<- as.character(dat2$`Admission Date`)
dat2$`Discharge Date` <-as.character(dat2$`Discharge Date`)
# 1) fix the admission date
# for the (same discharge date + same patient ID + same classification + same body region + same outcom
dat2 <-dat2 %>% group_by(`ROMS ID`,Outcome,`Body Region`,Classification, `Discharge Date`) %>% mutate(`A
# 2) fix the discharge date
# for the (same admission date + same patient ID + same classification + same body region + same outcom
dat2 <-dat2 %>% group_by(`ROMS ID`,Outcome,`Body Region`,Classification, `Admission Date`) %>% mutate(`D
```

```
# fix the age
dat2$Age <- floor(dat2$Age)

# fix the typo for outcomes
dat2$Outcome[dat2$Outcome =="Neck DISABILITY INDEX" ] <-  "NECK DISABILITY INDEX"
dat2$Outcome[dat2$Outcome =="neck DISABILITY INDEX" ] <-  "NECK DISABILITY INDEX"

# fix the typo for `chronic pain`
dat2$`Chronic Pain (Yes/No)`[dat2$`Chronic Pain (Yes/No)` =="yes"] <- "Yes"
dat2$`Chronic Pain (Yes/No)`[dat2$`Chronic Pain (Yes/No)` =="no"] <- "No"
dat2$`Chronic Pain (Yes/No)`[dat2$`Chronic Pain (Yes/No)` ==1] <- "Unknown"

# fix the typo for body regions
```

```
dat2$`Body Region`[dat2$`Body Region` == "knee"] <- "Knee"
dat2$`Body Region`[dat2$`Body Region` == "lumbar"] <- "Lumbar"

# remove the duplicated rows and select some columns.
#dat2 <- dat2 %>% distinct()

require(janitor)
dat2$`Injury Date` <- as.numeric(as.character(dat2$`Injury Date`))
```

```
## Warning: NAs introduced by coercion
```

```
dat2$`Injury Date` <- janitor::excel_numeric_to_date(dat2$`Injury Date`, date_system = "modern")
dat2$`Surgery Date` <- as.numeric(as.character(dat2$`Surgery Date`))
```

```
## Warning: NAs introduced by coercion
```

```
dat2$`Surgery Date` <- janitor::excel_numeric_to_date(dat2$`Surgery Date`, date_system = "modern")
```

```
#### fix the inconsistency of beg and final score
# 1) after grouping, if for one therapy the admission pain score is not consistent, use the first visit
dat2 <- dat2 %>%group_by(`ROMS ID`,Outcome,`Body Region`, Classification,`Admission Date`,`Discharge Da
# 2) after grouping, if for one therapy the admission outcome score is not consistent, use the first vi
dat2<- dat2 %>%group_by(`ROMS ID`,Outcome,`Body Region`,  Classification,`Admission Date`,`Discharge Da
# 3) after grouping, if for one therapy the discharge pain score is not consistent, use the last visit
dat2<- dat2 %>%group_by(`ROMS ID`,Outcome, `Body Region`, Classification,`Admission Date`,`Discharge Da
# 4) after grouping, if for one therapy the discharge outcome score is not consistent, use the last vis
dat2<- dat2 %>%group_by(`ROMS ID`,Outcome, `Body Region`, Classification,`Admission Date`,`Discharge Da

# fix the injury date, take the earliest.
# 1) in general
dat2$`Injury Date`<- as.character(dat2$`Injury Date`)
dat2 <- dat2 %>%group_by(`ROMS ID`,Outcome,`Body Region`, Classification,`Admission Date`,`Discharge Da
# 2) fix ID 882's injury date
dat2 <- dat2 %>%group_by(`ROMS ID`,`Body Region`, Classification,`Admission Date`,`Discharge Date`)%>%
 mutate(`Injury Date` = ifelse( length(unique(`Injury Date`))>=2, (min(`Injury Date`) ), `Injury Date`
```

```
# evaluate the clinical function for the pain level. if the pain in the last decreases at least 2 or fr
da_evaluate <- dat2  %>% mutate(pain_effect = ifelse((`Discharge Pain Score`-`Admission Pain`) <= -2 &

# evaluate the treatment effectiveness overall.
# 1)  1. for the outcome type =LOWER EXTREMITY FUNC SCALE, it's good discharge outcome score if `improv
#     2. if either the admission or outcome discharge score greater than 80 (the standard scale for thi
da_evaluate1<- da_evaluate %>%  filter(Outcome == "LOWER EXTREMITY FUNC SCALE") %>% mutate(effect_all =
da_evaluate1 <- da_evaluate1 %>% filter(`Admission Outcome Score`<=80) %>% filter(`Discharge Outcome Sco

# 2) deal with the outcome type =KNEE OUTCOME SURVEY, good if improve >= 9 or improve from >91 to 100,
da_evaluate2<- da_evaluate %>%  filter(Outcome == "KNEE OUTCOME SURVEY") %>% mutate(effect_all = ifelse
da_evaluate2 <- da_evaluate2 %>% filter(`Admission Outcome Score`<= 100) %>% filter(`Discharge Outcome S

# 3)check the outcome type in "MODIFIED LOW BACK DISABILITY QUESTIONNAIRE","Quick DASH","NECK DISABILIT
# (good if decreases more than 10, or decreases from less than 10 to 0, otherwise, bad.)
da_evaluate3 <- da_evaluate %>%  filter(Outcome %in% c("MODIFIED LOW BACK DISABILITY QUESTIONNAIRE","Qu
da_evaluate3 <- da_evaluate3 %>% filter(`Admission Outcome Score`<= 100) %>% filter(`Discharge Outcome S

# combine date
```

```
da_eval <- rbind(da_evaluate1,da_evaluate2,da_evaluate3)

# delete the distinct
da_eval  <- da_eval %>% distinct()
```

```
# we consider some features
subset <- da_eval[,c(2,5,6,7,8,15,16,17,18,20,21,22,25,26,27,28,29,30,31,32,33,37,38)]
# 4263 therapy?
subset.1 <- subset %>% group_by(`ROMS ID`,Outcome,`Body Region`,Classification, `Admission Date`,`Admis

# detected some conflict within age again.
# fix the age typo
subset.1$Age <- ifelse(subset.1$`ROMS ID` ==2435, 64, subset.1$Age)
subset.1$Age <- ifelse(subset.1$`ROMS ID` ==3539, NA, subset.1$Age)
subset.1$Age <- ifelse(subset.1$`ROMS ID` ==3957, 33, subset.1$Age)

# fix the conflict in chronic pain (by using unknown)
subset.1$`Chronic Pain (Yes/No)` <- ifelse(subset.1$`ROMS ID` %in% c(2435,2920,1418,2739), "Unknown", su
subset.1$`Chronic Pain (Yes/No)` <- as.factor(subset.1$`Chronic Pain (Yes/No)`)

# gather features together in new dataframe
df <- subset.1 %>% select(`ROMS ID`,Age, Gender, Outcome,`Body Region`,Classification, `Admission Date`

##ck <- df %>% group_by(`ROMS ID`,Age, Gender, Outcome,`Body Region`,Classification)%>% filter(n_distin

df <- df %>% distinct()

# give score
df <- df %>% mutate(painresult = ifelse(pain_effect =="good" , 1, 0)) %>% mutate(result = ifelse(effect_

#  2896 ID correspond to only 1 therapy (same period of time, same outcome type, same region, same clas
subset.1 <- df %>% group_by(`ROMS ID`) %>% filter(n()==1)

# repeated situation - 1276 records correspond to multiple records, maybe different outcome, or differe
subset.2 <- df %>% group_by(`ROMS ID`) %>% filter(n()>1)
subset.2$painresult <- as.numeric(subset.2$painresult)
subset.2$result <- as.numeric(subset.2$result)
```

```
# for the ID with repeated therapy summaries, check if the cause was outcome type
check <- subset.2 %>% group_by(`ROMS ID`,`Body Region`,Classification, `Admission Date`,`Discharge Date
# 1) if the only cause is outcome type, we evaluate the therapy on average.
check.1 <-check %>% filter(rep >=2 ) %>% mutate(painresult = ave(painresult),  result = ave(result),vis
check.1 <- check.1[, -c(23)]
#ck1<- check.1 %>% group_by(`ROMS ID`,Age, Gender, `Body Region`,Classification) %>% filter(n_distinct(
# also when we evaluate them by their average, some information becomes not valid logically.
check.1[,c(4,8,9, 11,12,13,14)] <- NULL
check.1 <- distinct(check.1)

# 2) if the cause is not outcome type, the repeated therapy summary can imply that 2 therapy exist, whe
check.2 <- check %>% filter(rep ==1 ) %>% mutate(pain = ave(painresult),  effectiveness = ave(result),vi
check.2 <- check.2[, -c(23,24,25)]
```

```
# combine the result, 3105 therapy in total
df_unique <- rbind(subset.1, check.2)
```

```
# combine the result including the repeated summaries of therapy, 3652 therapy in total
df_all <- dplyr::bind_rows(df_unique, check.1)

# remove the suspecious discharge date for id 2131, which discharge date is in June,2020
df_unique <- df_unique %>% filter(`ROMS ID` != 2131)
df_all <-df_all %>% filter(`ROMS ID` != 2131)
```

```
length(df_unique$`ROMS ID`)
```

```
## [1] 3104
```

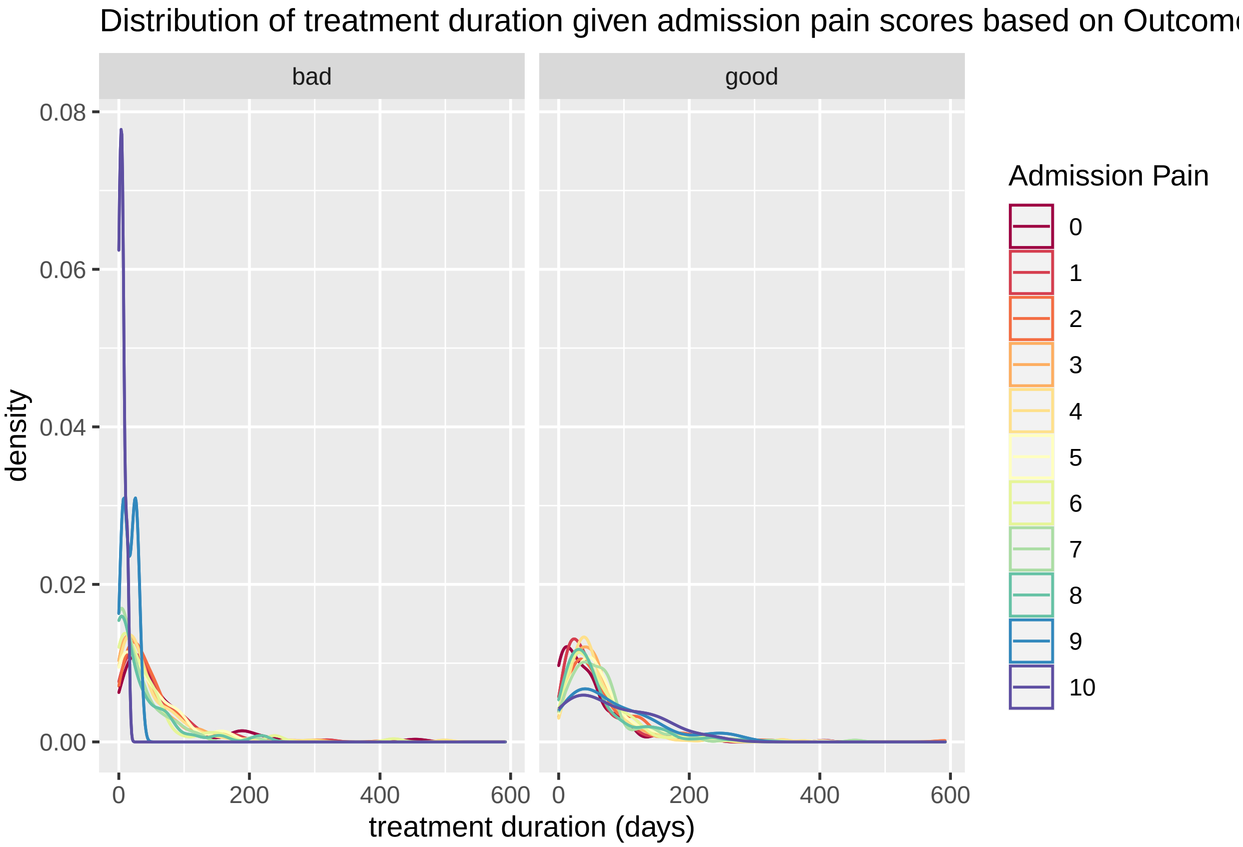
```
length(df_all$`ROMS ID`)
```

```
## [1] 3651
```

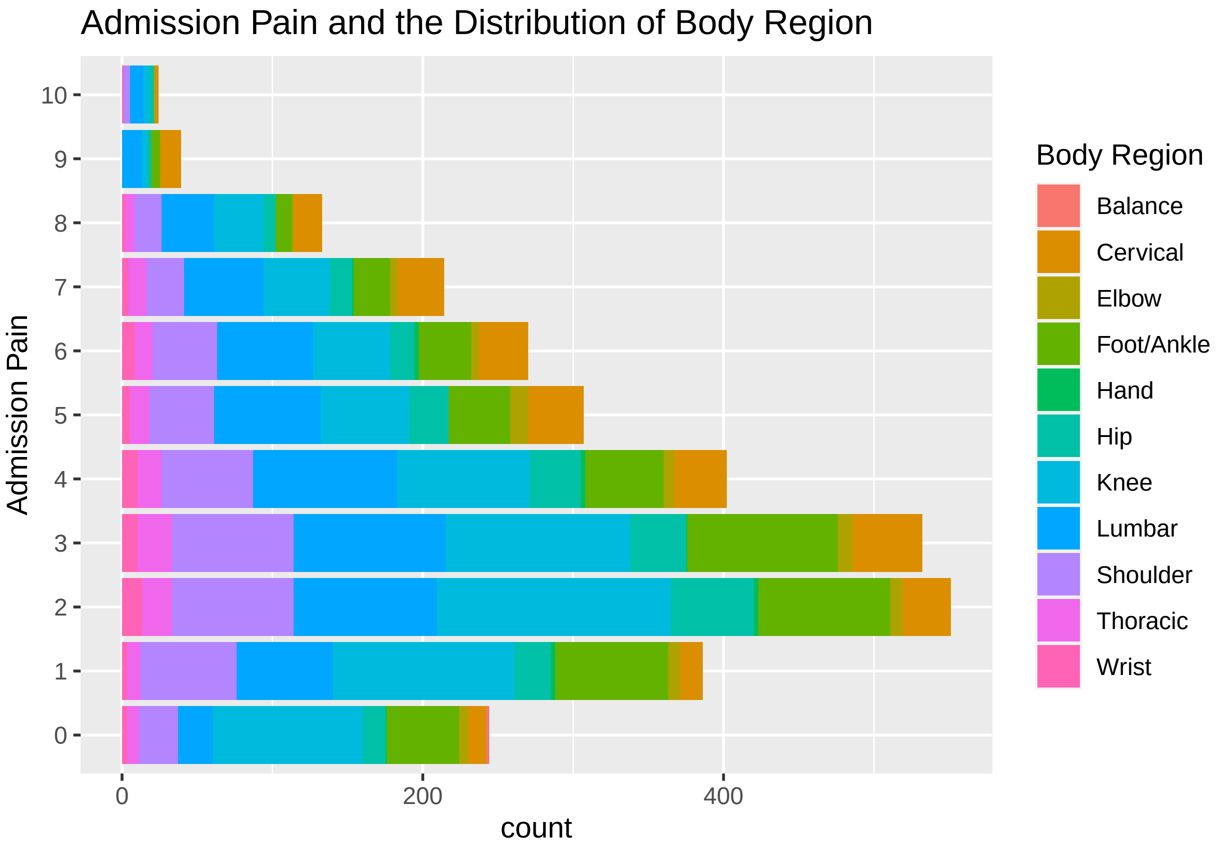
```
data <- df_unique
```

```
#da2$duration <- as.Date(da2$`Discharge Date`) - as.Date(da2$`Admission Date`)
data$`Admission Pain` <- as.factor(data$`Admission Pain`)
data <- data %>% drop_na(pain_effect)
ggplot(subset(data, `ROMS ID` != 2537 & 1893), aes(x=duration, color=`Admission Pain`)) +
    geom_density()+
    geom_line(stat="density")+
    scale_colour_brewer(palette = "Spectral")+
    facet_wrap(pain_effect~.)+
    labs( x=("treatment duration (days)"), y = ("density"), title= ("Distribution of treatment duration
    theme(plot.title = element_text(size =12))
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```

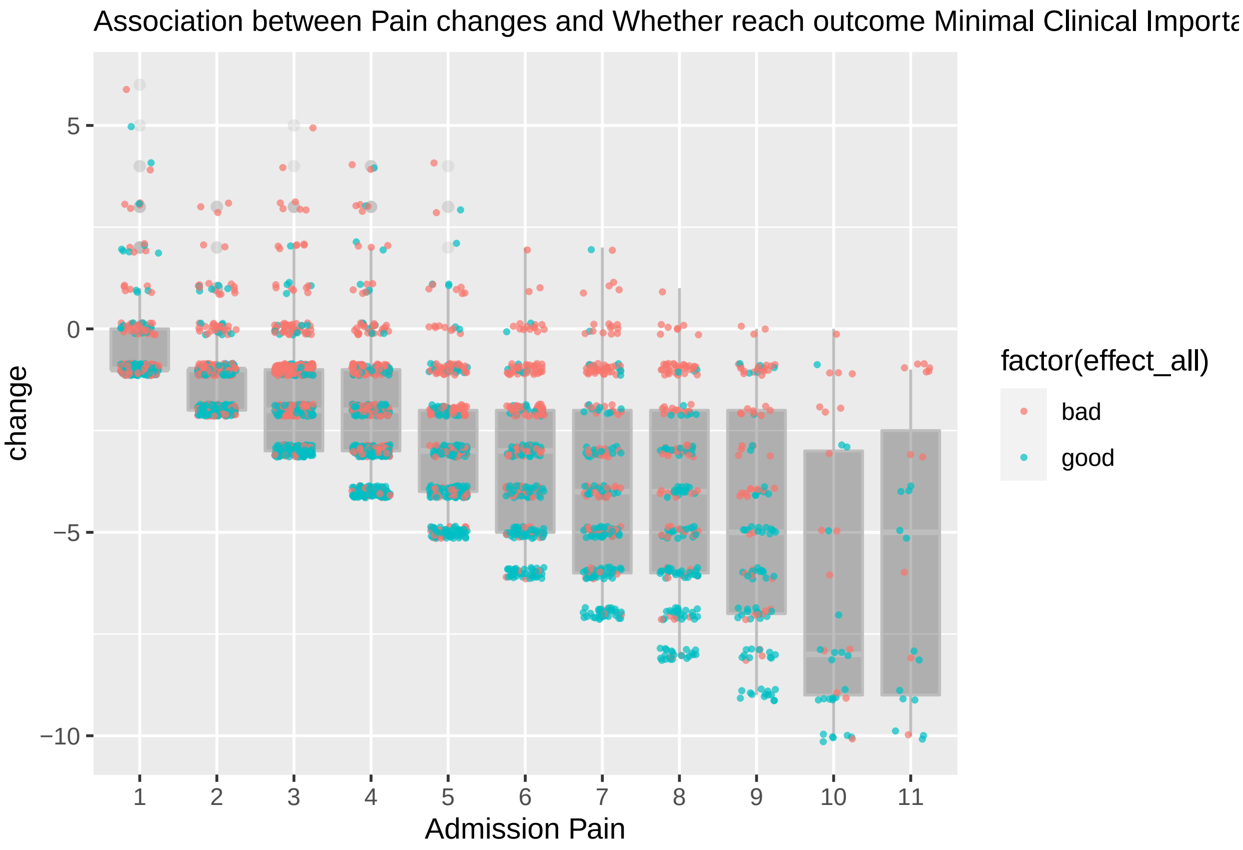


```r
ggplot(data, mapping = aes(x = `Admission Pain`, fill = `Body Region`)) +
    geom_bar(position="stack")+coord_flip() + labs(title = ("Admission Pain and the Distribution of Bod
```



```r
# pain changes scores vs admission pain scores. green color means the outcome improvement reaches the o
# pain relief and whether the outcome reaches minimal clinical important difference are positively asso
data$`Discharge Pain Score` <- as.numeric(data$`Discharge Pain Score`)
data$`Admission Pain` <- as.numeric(data$`Admission Pain`)
data2 <- data %>% mutate (change = `Discharge Pain Score` - `Admission Pain`)
data2$`Admission Pain`<-as.factor(data2$`Admission Pain`)

ggplot(data2, aes(x = `Admission Pain`, y = change))+ geom_boxplot(col = "gray", alpha = 0.3,fill= "gra
    theme(plot.title = element_text(size =11))
```


Association between Pain changes and Whether reach outcome Minimal Clinical Importa
change
5
0
−5
−10
1
2
3
4
5
6
7
8
9
10
11
Admission Pain
factor(effect_all)
bad
good