Study on the Returning rate for Patients with Lung Nodule. Part 1.

Jingjian Gao, Yaquan Yang, Jingyao Wang, Jin Yan

2022-12-10

Introduction

The project is looking at patient information from a lung cancer screening database. BMC has a lung cancer screening program for individuals that meet specific criteria (amount of tobacco use & duration of smoking). Depending on the findings, people are screened with low-dose CT scans and are recommended to have a repeat scan within a certain period. We tried using different graphing methods to show the relationship between the number of follow-up scans and potential factors from the data we got. Due to the limited time, we have performed EDA on the data, which is shown below.

Project description

We have 267 patient records of the lung cancer screening program at Boston Medical Center. We use the Lung-RADS score, which is a radiologic grading system to determine how urgently someone should get repeat imaging.

- Rads-3 1-2%
- Rads-4a 5-15%
- Rads-4b >15%

These patients may not have been diagnosed with lung cancer yet, but the initial CT scan indicated that there are lung nodules present. The larger the nodules, the higher the chance of developing lung cancer. Our client wants to find out what factors are associated with obtaining any follow-up CT scanning of Lung Rads 3-4 findings. He also wants to investigate whether reminder letters sent by lung cancer screening programs have increased follow-up or timely follow-up scans; this requires further information than the data we have currently, hence we will not address it now.

Loading Required Package and import Data

```
library(data.table)
library(readxl)
library(ggplot2)
library(plotly)
library(ggalluvial)
library(stringr)
library(ggridges)
library(patchwork)
library(flextable)
library(gridExtra)
```

```r
info<-read_xlsx('De-identified Data.xlsx')
setDT(info)
info<-info[order(Identifier)]
```

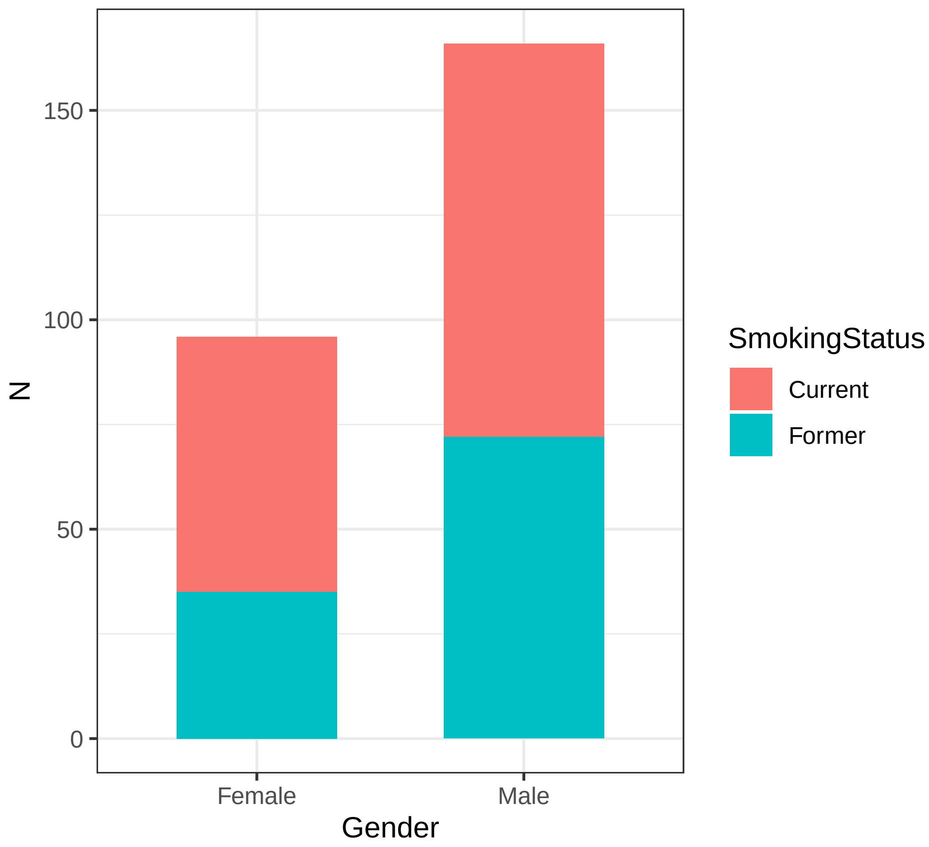
Gender & SmokingStatus

From this graph we can see that even after being diagnosed as having lung cancers, most of the patients continues to smoke.

```r
info[,Gender:=str_to_title(Gender)]
info[!Gender %in% c('Male','Female'),Gender:=NA]

info[,SmokingStatus:=str_to_title(SmokingStatus)]
info[SmokingStatus=='Unknown',SmokingStatus:=NA]


plot2.data<-info[,.N,.(Gender,SmokingStatus)]
plot2.data<-plot2.data[!is.na(Gender) & !is.na(SmokingStatus)]

flextable(plot2.data) %>% theme_vanilla()
```

Gender	SmokingStatus	N
Male	Current	94
Female	Current	61
Female	Former	35
Male	Former	72

```r
ggplot(plot2.data,aes(x=Gender,y=N,fill=SmokingStatus))+
      geom_bar(stat='identity',position='stack',width=0.6)+
      theme_bw()
```

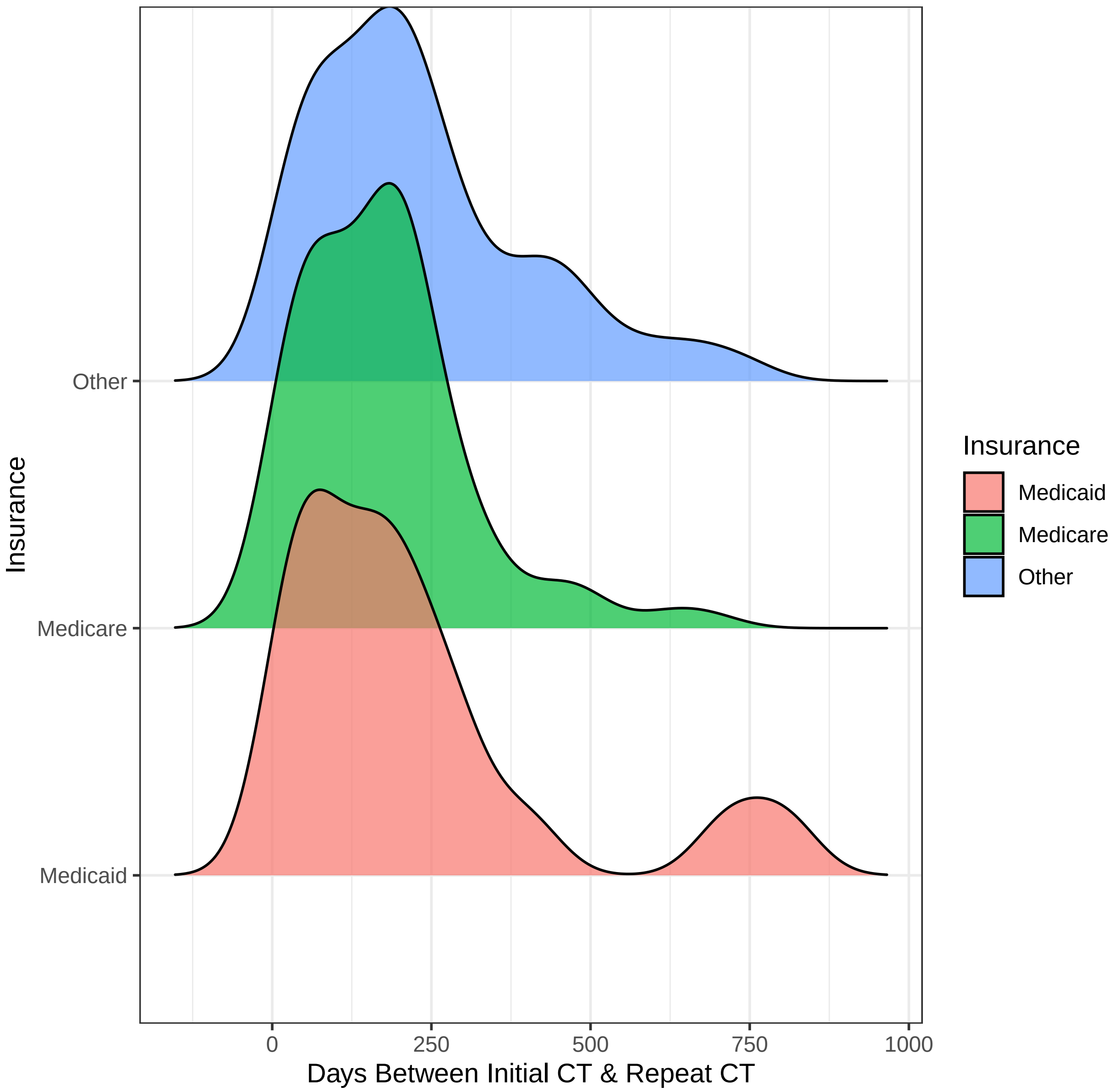


Days Between Initial CT & Repeat CT & Insurance

From the Density Ridge Graph, we can see that there is a possibility that the follow up letters are sent around 550 days for Medicaid group.

```
info[!Insurance %in% c('Medicare','Medicaid'),Insurance:='Other']

ggplot(info,aes(x=`Days Between Initial CT & Repeat CT`,y=Insurance,fill=Insurance))+
   geom_density_ridges(alpha=0.7)+
   theme_bw()
```

```
## Picking joint bandwidth of 53.5
```

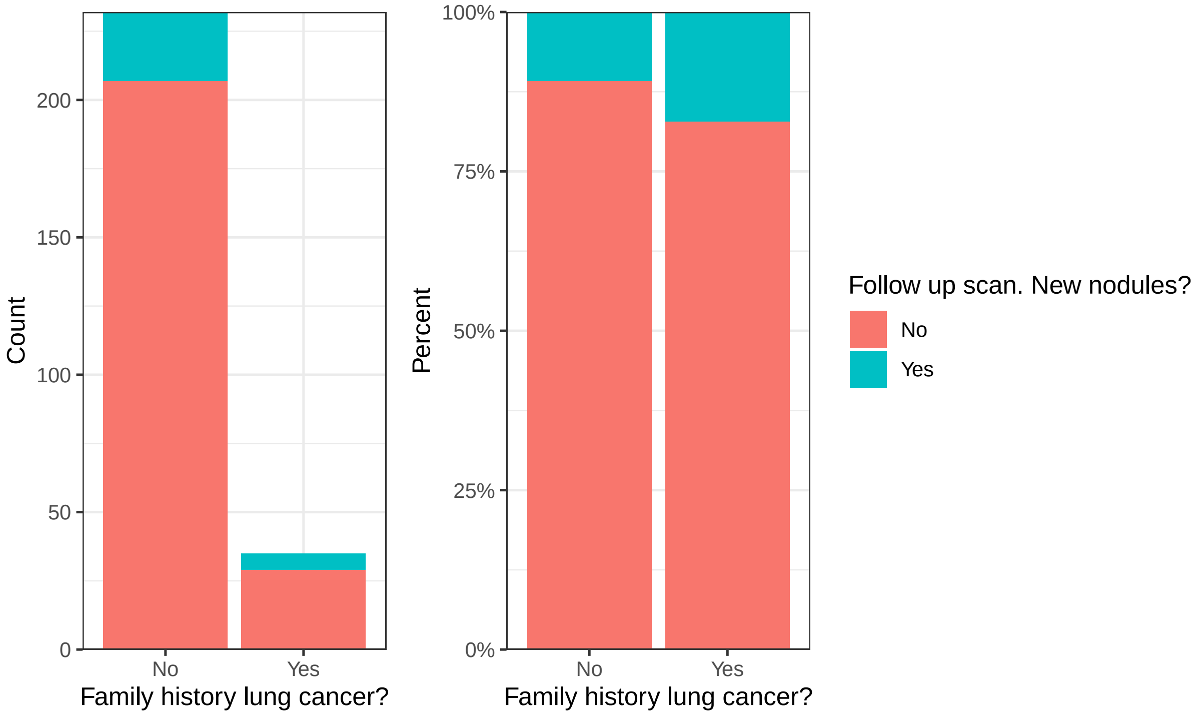


Family history lung cancer? & Follow up scan. New nodules?

```
info[`Family history lung cancer?`=='none',`Family history lung cancer?`:='no']
info[,`Family history lung cancer?`:=str_to_title(`Family history lung cancer?`)]

plot5.data<-info[,.N,.(`Family history lung cancer?`,`Follow up scan. New nodules?`)]

flextable(plot5.data) %>% theme_vanilla()
```

Family history lung cancer?	Follow up scan. New nodules?	N
No	No	207
No	Yes	25
Yes	No	29
Yes	Yes	6

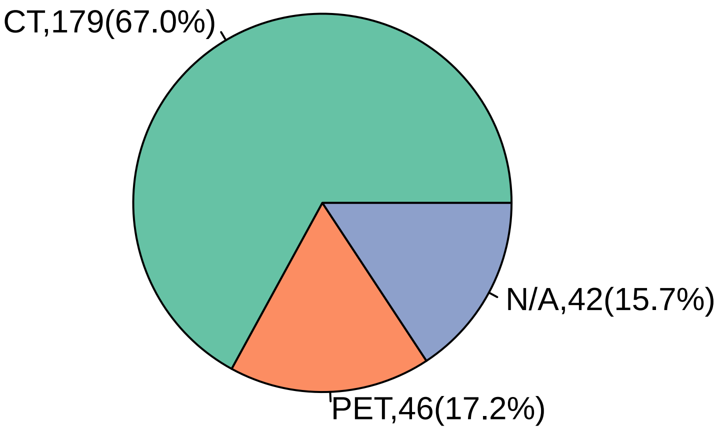
```r
p52<-ggplot(plot5.data,
           aes(x=`Family history lung cancer?`,y=N,fill=`Follow up scan. New nodules?`))+
           geom_bar(stat='identity',position=position_fill(reverse=T))+
           theme_bw()+
           scale_y_continuous(label=scales::percent,expand=expansion(0))+
           labs(y='Percent')


p51<-ggplot(plot5.data,
      aes(x=`Family history lung cancer?`,y=N,fill=`Follow up scan. New nodules?`))+
      geom_bar(stat='identity',position=position_stack(reverse=T))+
      theme_bw()+
      scale_y_continuous(expand=expansion(c(0)))+
      labs(y='Count')


p51+p52+plot_layout(guides='collect')
```



The graph shows that most of the people would not go to follow ups.

What type of follow up scan was done?

```
library(RColorBrewer)
library(scales)
info[,`What type of follow up scan was done? (CT vs. PET)`:=
       str_to_upper(`What type of follow up scan was done? (CT vs. PET)`)]

plot6.data<-info[,.N,`What type of follow up scan was done? (CT vs. PET)`]
pie(plot6.data$N,paste(plot6.data[[1]],paste0(plot6.data$N,"(",percent(plot6.data$N/sum(plot6.data$N)),
```

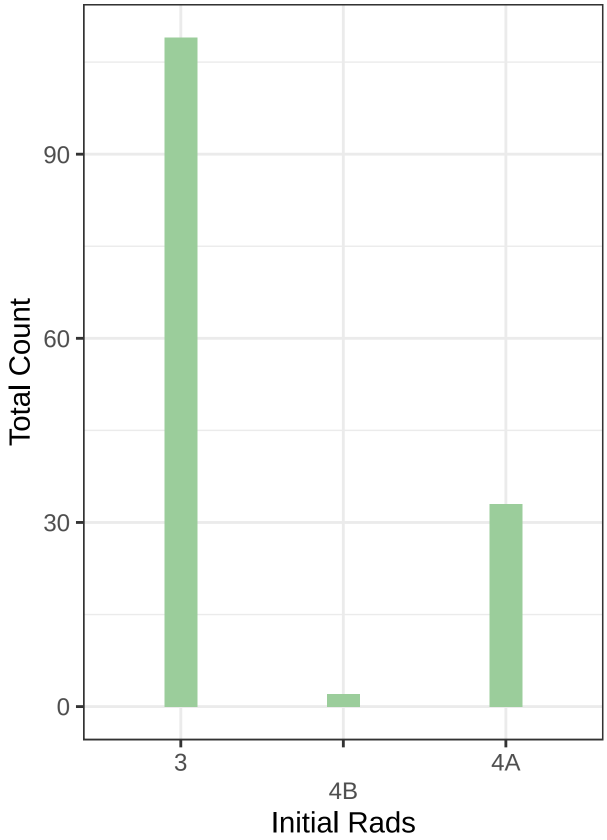


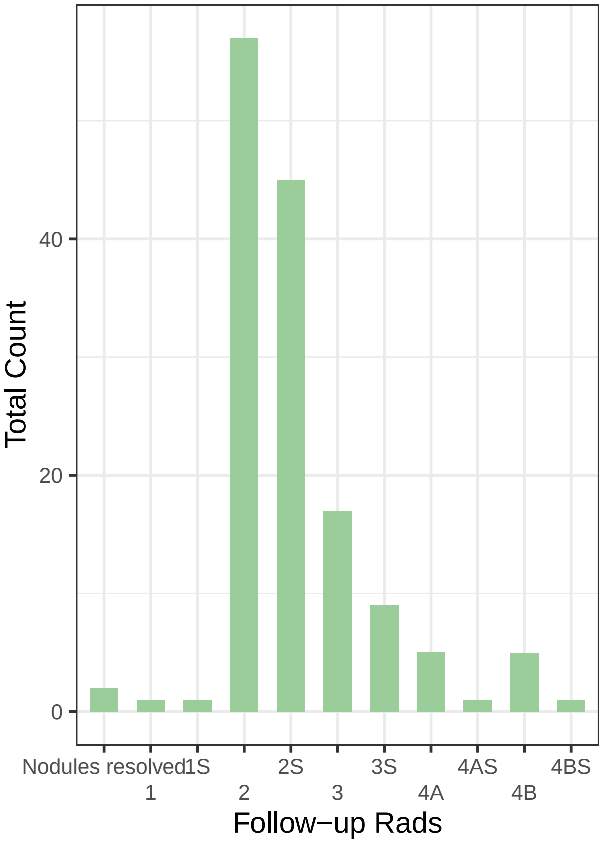
Within the group of people who did the follow up scans, 67% of the people did CT scans, PET 17.2%

Rads change after Followup.

```
plot1.data<-info[,.(Rads.Initial=`Rads Initial`, Rads.followup=`Rads Category Follow up scan`)]

plot1.data[Rads.followup %in% c('N/A',
          'No comment of rads category on follow up scan',
          'not specified on follow up scan',
          'Follow up scan did not specify'), Rads.followup:='N/A']
```

```
plot1.data[Rads.followup!='N/A']->plot1.data
plot1.data[Rads.followup!='PET']->plot1.data

plot1.data<-plot1.data[,.N,.(Rads.Initial, Rads.followup)]

plot1.data[,Rads.Initial:=ordered(Rads.Initial,
                                  levels=c('3','4B','4A','4X'),
                                  labels=c('3','4B','4A','4X'))]

plot1.data[,Rads.followup:=ordered(Rads.followup,
                                  levels=c("Nodules resolved",
                                  "1" , "1S" , "2" ,  "2S",   "3"  ,   "3S" ,
                                  "4A" , "4AS" ,    "4B" ,   "4BS" ,
                                  "PET" ,  "N/A"))]

plot1.data<-plot1.data[order(Rads.Initial,Rads.followup)]


p1 <- ggplot(plot1.data,aes(x=Rads.Initial, y=plot1.data$N))+
      geom_bar(stat='identity',fill='darkseagreen3',width=0.2)+
      theme_bw()+
scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
    labs(x='Initial Rads',
     y='Total Count')

p2 <- ggplot(plot1.data,aes(x=Rads.followup, y=plot1.data$N))+
      geom_bar(stat='identity',fill='darkseagreen3',width=0.6)+
      theme_bw()+
    theme(axis.text=element_text(size=8))+
scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
    labs(x='Follow-up Rads',
     y='Total Count')

grid.arrange(p1,p2,ncol = 2)
```





We also made graphs showing the Rads scores for the initial and follow-up scans. The Rads score reflects the likelihood of the nodule developing into cancer. The higher the rads score, the higher the likelihood of developing cancer. As we can see from the graph, there has been a significant decrease in the number of level 3 and level 4 patients. Most of the patients' diseases progressed in a good direction after the follow-up scan was performed. This indicates the importance of timely follow-up scans to monitor the disease and active treatment.

Conclusion

We demonstrated some associations between the factors mentioned in the data and the number of follow-up checks. The smoking history seems to be an essential factor to consider in future analysis. All 267 patients from the data had some kind of smoking history. From the Ridge graph, we speculate that the reminder letter may have been sent around day 550. In future analyses, we can do a more detailed EDA on the patients' demographic information.