NIH Clinical Center Chest X-Rays Data Analysis

Serena Gibbons

I. Introduction

In this project I would like to show a multi-panel plot of the 14 common thorax disease categories included in the NIH Chest X-Ray Dataset obtained from https://nihcc.app.box.com/v/ChestXray-NIHCC. I would like to use this dataset to plot visualizations of patient population distributions and their clinical diagnoses, including the frequency of the 14 disease labels distributed among gender and age.

II. The Data

In this dataset, chest X-Ray images are classified as one of 15 classes (14 diseases, and "No findings"). The dataset was extracted from the clinical picture archiving and communication system (PACS) database at National Institutes of Health Clinical Center and consists of approximately 60% of all frontal chest x-rays in the hospital. The dataset consists of 112,120 fontal view x-ray images of 30,805 unique patients with the 14 disease image labels (where each image can have multiple labels). This data was collected with the goal to achieve better clinically relevant computer aided detection and diagnosis (CAD) of chest x-rays in medical centers. I am interested in using this data to observe the distribution of different patient populations among the 14 pathologies.

```
df <- read.csv("Data_Entry_2017.csv")
head(df)
```

```
##        Image.Index         Finding.Labels Follow.up.. Patient.ID Patient.Age
## 1 00000001_000.png           Cardiomegaly           0          1          58
## 2 00000001_001.png Cardiomegaly|Emphysema           1          1          58
## 3 00000001_002.png  Cardiomegaly|Effusion           2          1          58
## 4 00000002_000.png             No Finding           0          2          81
## 5 00000003_000.png                 Hernia           0          3          81
## 6 00000003_001.png                 Hernia           1          3          74
##   Patient.Gender View.Position OriginalImage.Width Height.
## 1              M            PA                2682    2749
## 2              M            PA                2894    2729
## 3              M            PA                2500    2048
## 4              M            PA                2500    2048
## 5              F            PA                2582    2991
## 6              F            PA                2500    2048
##   OriginalImagePixelSpacing.x    y.
## 1                       0.143 0.143
## 2                       0.143 0.143
## 3                       0.168 0.168
## 4                       0.171 0.171
## 5                       0.143 0.143
## 6                       0.168 0.168
```

```r
summary(df)
```

```
##  Image.Index        Finding.Labels      Follow.up..        Patient.ID
##  Length:112120      Length:112120      Min.   :  0.000   Min.   :    1
##  Class :character   Class :character   1st Qu.:  0.000   1st Qu.: 7311
##  Mode  :character   Mode  :character   Median :  3.000   Median :13993
##                                        Mean   :  8.574   Mean   :14346
##                                        3rd Qu.: 10.000   3rd Qu.:20673
##                                        Max.   :183.000   Max.   :30805
##   Patient.Age    Patient.Gender     View.Position      OriginalImage.Width
##  Min.   :  1.0   Length:112120      Length:112120      Min.   :1143
##  1st Qu.: 35.0   Class :character   Class :character   1st Qu.:2500
##  Median : 49.0   Mode  :character   Mode  :character   Median :2518
##  Mean   : 46.9                                         Mean   :2646
##  3rd Qu.: 59.0                                         3rd Qu.:2992
##  Max.   :414.0                                         Max.   :3827
##     Height.     OriginalImagePixelSpacing.x       y.
##  Min.   : 966   Min.   :0.1150              Min.   :0.1150
##  1st Qu.:2048   1st Qu.:0.1430              1st Qu.:0.1430
##  Median :2544   Median :0.1430              Median :0.1430
##  Mean   :2486   Mean   :0.1556              Mean   :0.1556
##  3rd Qu.:2991   3rd Qu.:0.1680              3rd Qu.:0.1680
##  Max.   :4715   Max.   :0.1988              Max.   :0.1988
```

III. Data Cleaning and Manipulation

There are a few values in Patient Age that do not make sense (ranging from 148 to 414). These could possibly be age in months, weeks, or days but it is impossible to know for sure so these rows will be removed with a cut off of 100. I will also remove columns that will not be used (such as image index and columsn pertaining to image dimensions).

```r
mydata <- df[-(which(df$Patient.Age > 100)), ]  # Remove rows with Patient Age greater than 100
mydata <- mydata[, c(2:6)]  # Remove unused columns
head(mydata)
```

```
##           Finding.Labels Follow.up.. Patient.ID Patient.Age Patient.Gender
## 1           Cardiomegaly           0          1          58              M
## 2 Cardiomegaly|Emphysema           1          1          58              M
## 3  Cardiomegaly|Effusion           2          1          58              M
## 4             No Finding           0          2          81              M
## 5                 Hernia           0          3          81              F
## 6                 Hernia           1          3          74              F
```

To analzye the distributions of single diagnoses without comorbiditie, I created a new data frame without rows of multiple diagnoses in the Finding Labels column (e.g. “Cardiomegaly|Effusion”).

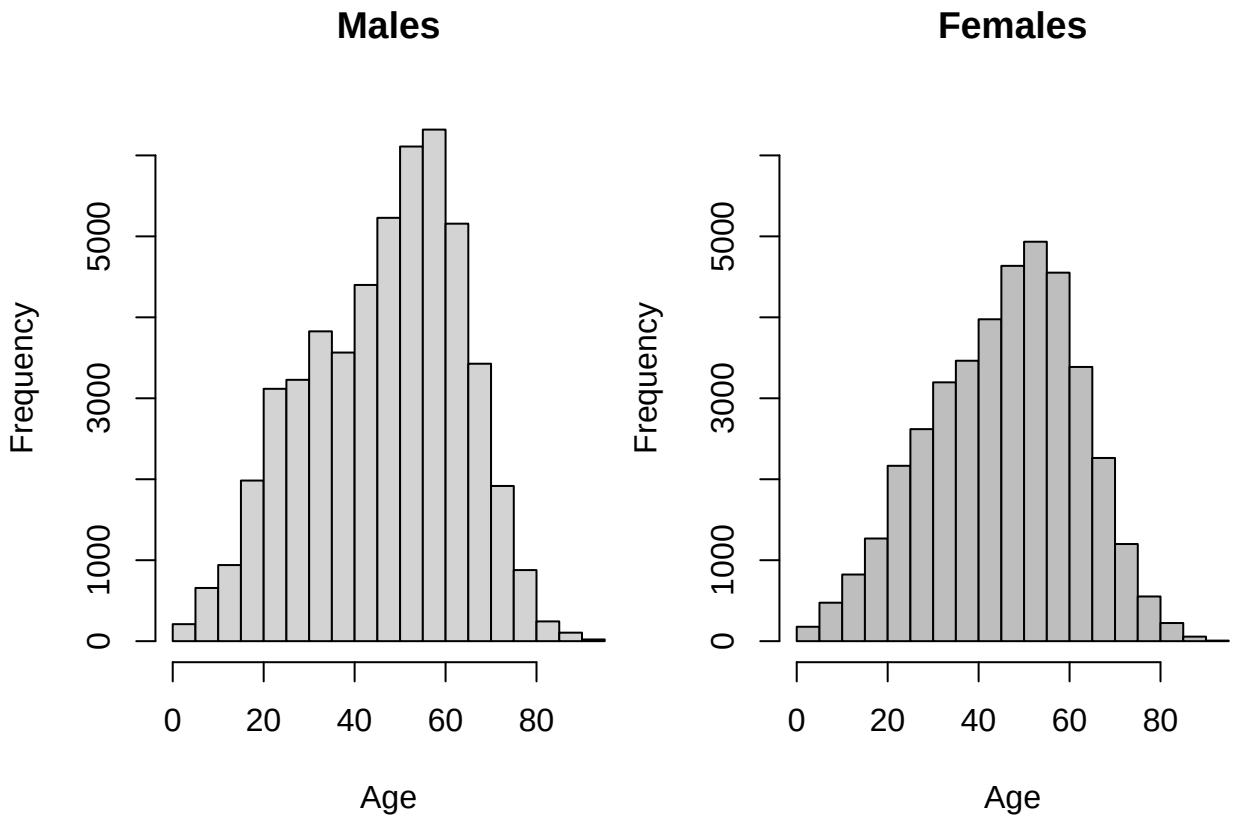
```r
# create new data frame of single diagnoses
mydata_single <- mydata[with(mydata, which(Finding.Labels == "Cardiomegaly" |
    Finding.Labels == "Emphysema" | Finding.Labels == "Effusion" | Finding.Labels ==
    "Hernia" | Finding.Labels == "Nodule" | Finding.Labels == "Atelectasis" |
    Finding.Labels == "Pleural_Thickening" | Finding.Labels == "Mass" | Finding.Labels ==
```

```
    "Edema" | Finding.Labels == "Consolidation" | Finding.Labels == "Infiltration" |
    Finding.Labels == "Fibrosis" | Finding.Labels == "Pneumonia" | Finding.Labels ==
    "Pneumothorax" | Finding.Labels == "No Finding")), ]
```

IV. Patient population distributions

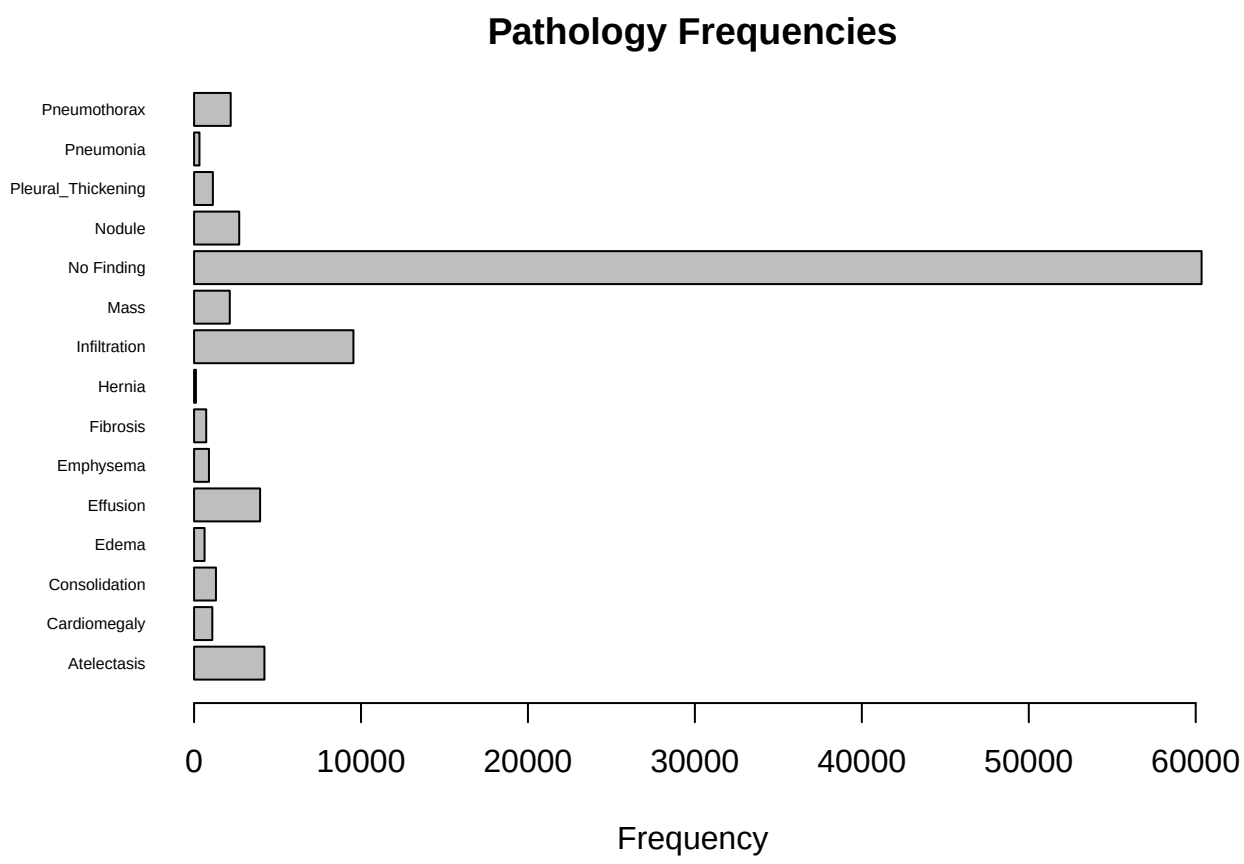
I will create 3 different types of visualizations: age distributions of chest X-rays by patient gender, frequncies of pathologies, and pathologies among patient ages.

```
x11(title = "Age distributions by gender", height = 6, width = 4)
layout(matrix(c(1, 2), nrow = 1))  # create 2 panel plotting windows
par(mar = c(4.1, 4.1, 3.6, 0))
mydata_singleM <- mydata_single[which(mydata_single$Patient.Gender == "M"),
    ]
mydata_singleF <- mydata_single[which(mydata_single$Patient.Gender == "F"),
    ]
hist(mydata_singleM$Patient.Age, xlab = "Age", main = "Males", ylim = c(0, 6500))
hist(mydata_singleF$Patient.Age, col = "grey", xlab = "Age", main = "Females",
    ylim = c(0, 6500))
```

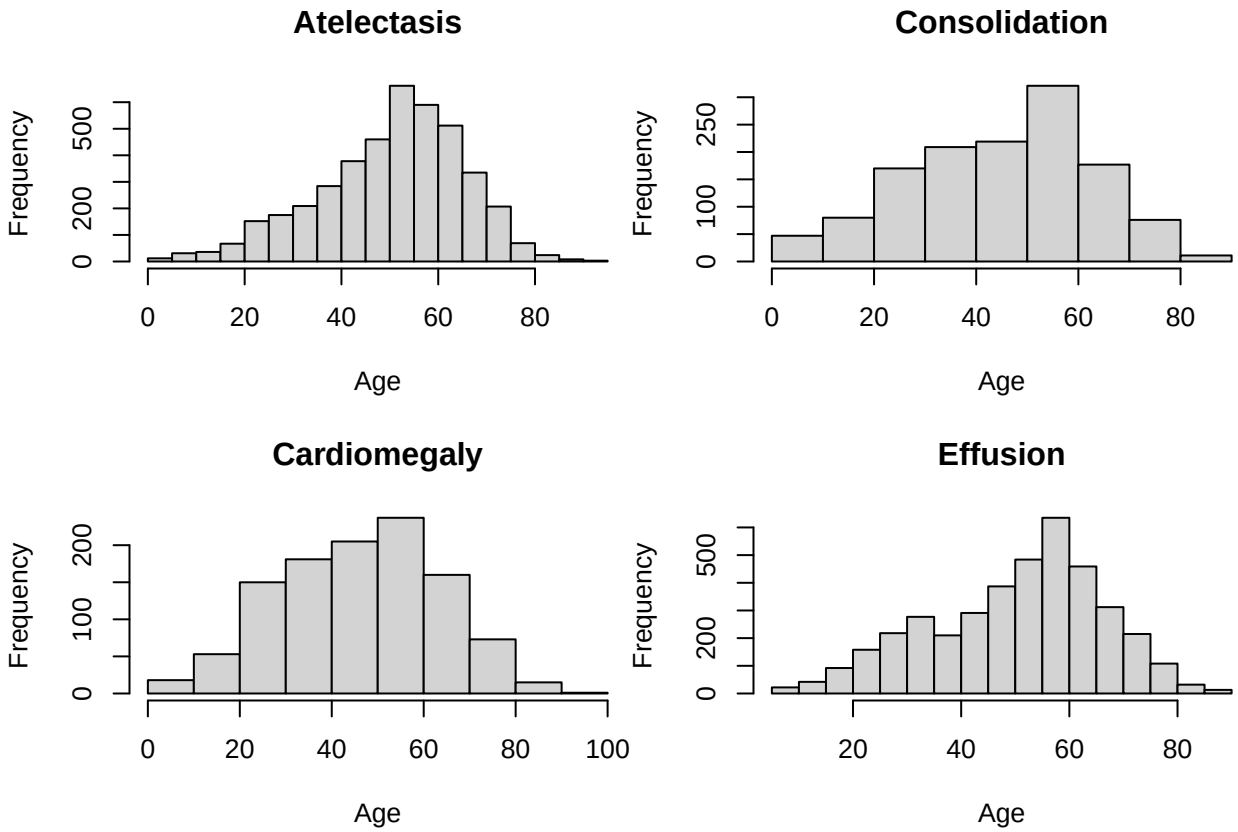



```
diseaseFreq <- as.data.frame(table(mydata$Finding.Labels))
names(diseaseFreq) <- c("Disease", "Freq")
singleDisease <- diseaseFreq[with(diseaseFreq, which(Disease == "Cardiomegaly" |
    Disease == "No Finding" | Disease == "Emphysema" | Disease == "Effusion" |
```

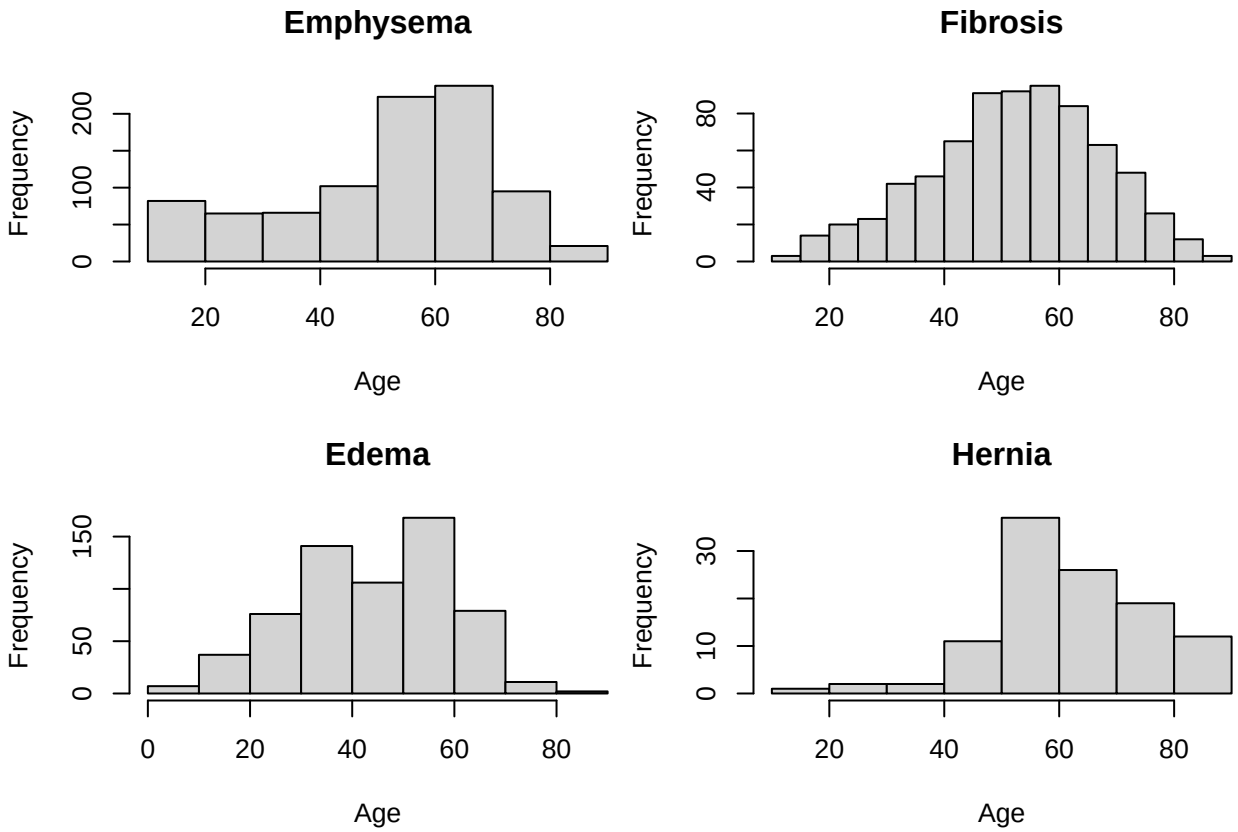
```
    Disease == "Hernia" | Disease == "Nodule" | Disease == "Atelectasis" | Disease ==
    "Pleural_Thickening" | Disease == "Mass" | Disease == "Edema" | Disease ==
    "Consolidation" | Disease == "Infiltration" | Disease == "Fibrosis" | Disease ==
    "Pneumonia" | Disease == "Pneumothorax")), ]
x11(title = "Age distributions by gender", height = 6, width = 4)
par(mar = c(4, 5, 2, 1))
barplot(singleDisease$Freq, names.arg = singleDisease$Disease, las = 1, cex.names = 0.5,
    horiz = TRUE, xlab = "Frequency", main = "Pathology Frequencies")
```



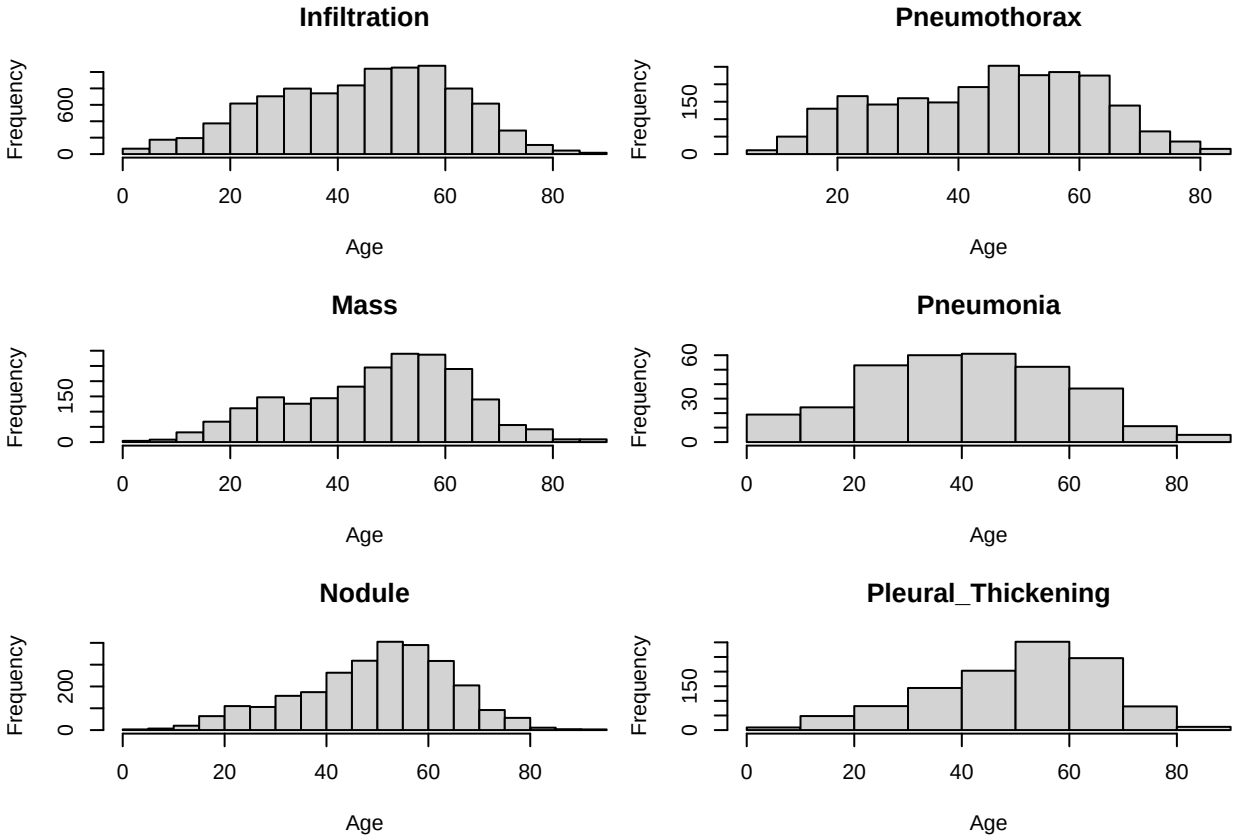

```
# Create histograms of each disease pathology among patient ages
x11(height = 7, width = 5)
layout(matrix(c(1:4), nrow = 2))  # create 4 panel plotting window
par(mar = c(4, 4, 3.6, 0))
hist(mydata_single[which(mydata_single$Finding.Labels == "Atelectasis"), ]$Patient.Age,
    xlab = "Age", main = "Atelectasis")
hist(mydata_single[which(mydata_single$Finding.Labels == "Cardiomegaly"), ]$Patient.Age,
    xlab = "Age", main = "Cardiomegaly")
hist(mydata_single[which(mydata_single$Finding.Labels == "Consolidation"), ]$Patient.Age,
    xlab = "Age", main = "Consolidation")
hist(mydata_single[which(mydata_single$Finding.Labels == "Effusion"), ]$Patient.Age,
    xlab = "Age", main = "Effusion")
```




```
x11(height = 7, width = 5)
layout(matrix(c(1:4), nrow = 2))  # Create 4 panel plotting window
par(mar = c(4, 4, 3.6, 0))
hist(mydata_single[which(mydata_single$Finding.Labels == "Emphysema"), ]$Patient.Age,
    xlab = "Age", main = "Emphysema")
hist(mydata_single[which(mydata_single$Finding.Labels == "Edema"), ]$Patient.Age,
    xlab = "Age", main = "Edema")
hist(mydata_single[which(mydata_single$Finding.Labels == "Fibrosis"), ]$Patient.Age,
    xlab = "Age", main = "Fibrosis")
hist(mydata_single[which(mydata_single$Finding.Labels == "Hernia"), ]$Patient.Age,
    xlab = "Age", main = "Hernia")
```




```
x11(height = 7, width = 5)
layout(matrix(c(1:6), nrow = 3))  # Create 6 panel plotting window
par(mar = c(4, 4, 3.6, 0))
hist(mydata_single[which(mydata_single$Finding.Labels == "Infiltration"), ]$Patient.Age,
    xlab = "Age", main = "Infiltration")
hist(mydata_single[which(mydata_single$Finding.Labels == "Mass"), ]$Patient.Age,
    xlab = "Age", main = "Mass")
hist(mydata_single[which(mydata_single$Finding.Labels == "Nodule"), ]$Patient.Age,
    xlab = "Age", main = "Nodule")
hist(mydata_single[which(mydata_single$Finding.Labels == "Pneumothorax"), ]$Patient.Age,
    xlab = "Age", main = "Pneumothorax")
hist(mydata_single[which(mydata_single$Finding.Labels == "Pneumonia"), ]$Patient.Age,
    xlab = "Age", main = "Pneumonia")
hist(mydata_single[which(mydata_single$Finding.Labels == "Pleural_Thickening"),
    ]$Patient.Age, xlab = "Age", main = "Pleural_Thickening")
```



Note: I had originally wanted to create a single 14-panel plot using layout(matrix(c(1:14), nrow = 7)) however I recieved the error message: "Error in plot.new() : figure margins too large" when plotting the histograms.

References:

NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community. (2017, September 27). Retrieved from https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. IEEE CVPR 2017, http://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf