Project Check in 1

2024-10-03

Below you will find the template for Project Check in 1. Each group member should submit a project check in document. Work with your team to ensure you limit duplicated work. The example code below is instructional and should not appear in your submission. Please do feel free to adapt it to your own dataset.

1. ***If your dataset has changed*** from what you submitted for Project Check in 0, please record:

- number of observations (rows)
- number of variables (columns)
- number of missing values
- names of particular columns of interest (if there are too many to print all of them!)
- data source and links to any accompanying documentation

```r
source <-
  read.csv("cervical-programme-annual-2022-23-csvs/kc61_sample_result_age_source.csv")

dim(source)
```

```
## [1] 330  17
```

```r
#columns of interest: Indicator (age), Borderline changes, mild, moderate, and severe dyskaryosis

#only column that has nas is severe dyskaryosis
source$Severe_dyskaryosis <- gsub("not included", NA, source$Severe_dyskaryosis)
sum(is.na(source$Severe_dyskaryosis) == TRUE)
```

```
## [1] 90
```

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

2. Show summary statistics for the five variables of the most interest.

```r
uniq_ind <- unique(source$Indicator)
for (i in 1:length(uniq_ind)){
  source$Indicator <- gsub(uniq_ind[i], i, source$Indicator)
}
source$Indicator <- as.numeric(source$Indicator)
#changing each age group to correspond with a number from 1-19. for example, <20 is 1

for (i in 1:length(uniq_ind)){
  df <- data.frame(source[source$Indicator == i, ])
  df_name <- paste0("df", i)
  assign(df_name, df)
}
#creating a new data frame for every age group and organizations data was taken from
#`assign` function taken and adapted from online source

for(i in 1:length(uniq_ind)){
  source_ind <- source[source$Indicator == (uniq_ind[i]), ]
  print(summary(source_ind$Borderline_changes))
}
#summary for borderline changes for each age group

for(i in 1:length(uniq_ind)){
  source_ind <- source[source$Indicator == (uniq_ind[i]), ]
  print(summary(source_ind$Mild_dyskaryosis))
}
#summary for mild dyskaryosis for each age group

for(i in 1:length(uniq_ind)){
  source_ind <- source[source$Indicator == (uniq_ind[i]), ]
  print(summary(source_ind$Moderate_dyskaryosis))
}
#summary for moderate dyskaryosis for each age group

#removing NAs and converting to numeric values
severe_NA <- source[is.na(source$Severe_dyskaryosis) == FALSE, ]
severe_NA$Severe_dyskaryosis <- as.numeric(severe_NA$Severe_dyskaryosis)

for(i in 1:length(uniq_ind)){
  severe_ind <- severe_NA[source$Indicator == (uniq_ind[i]), ]
  print(summary(severe_ind$Severe_dyskaryosis))
}
#summary for severe dyskaryosis for each age group
#summary output excluded due to large output of information.
```

- 1: <20 years old (non-inclusive)
- 2: 20-24 years old (inclusive)
- 3: 25-29 years old (inclusive)
- 4: 30-34 years old (inclusive)
- 5: 35-39 years old (inclusive)
- 6: 40-44 years old (inclusive)
- 7: 45-49 years old (inclusive)
- 8: 50-54 years old (inclusive)
- 9: 55-59 years old (inclusive)
- 10: 60-64 years old (inclusive)
- 11: 65-69 years old (inclusive)
- 12: 70-74 years old (inclusive)
- 13: >=75 years old (inclusive)
- 14: data taken from GP
- 15: data taken from NHSCC
- 16: data taken from GUM
- 17: data taken from NHS Hospital
- 18: data taken from Private sources

- 19: data taken from Other sources

3. Show another set of summary statistics by filtering on a column of interest. This could be years, teams, genres. Dig a little deeper here! Think about what might have a different distribution!

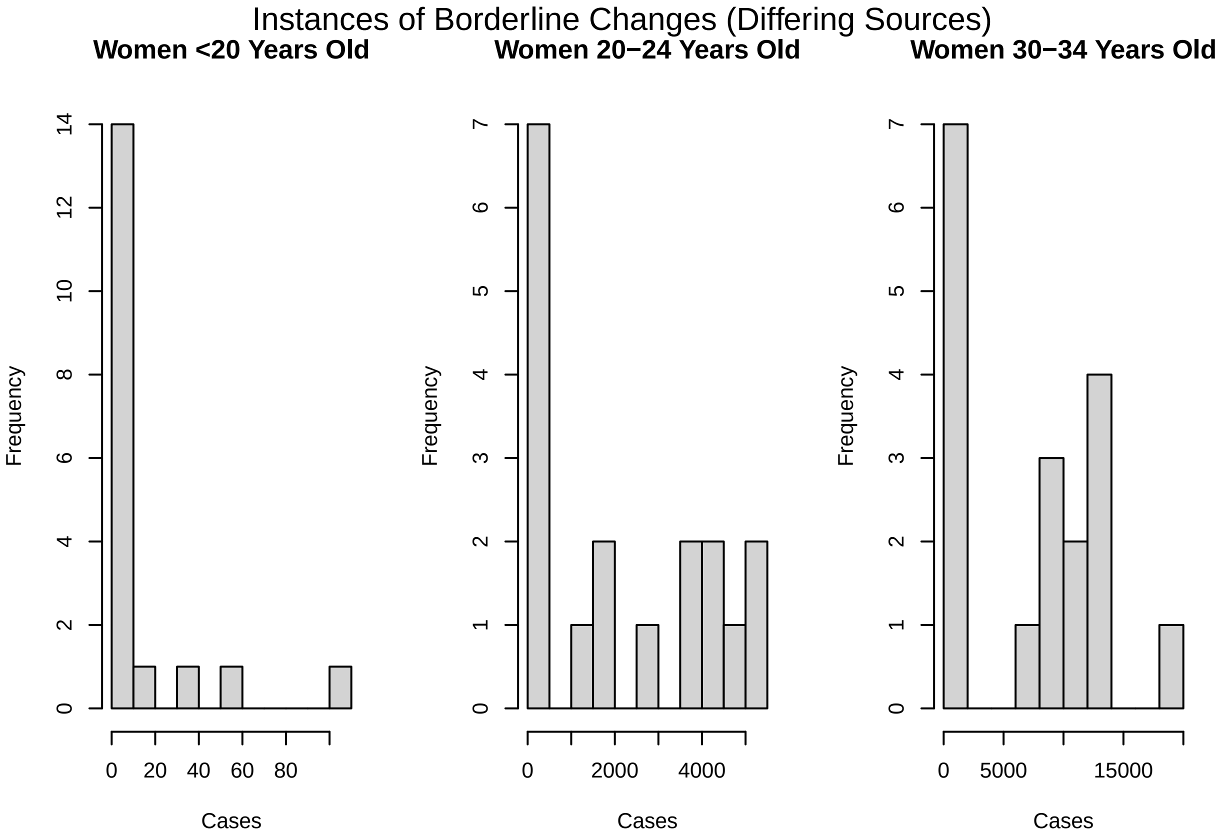
Filtered in #2 by age. See above

4. Visualize the distribution of at least three variables. For example, below I've made a histogram to investigate the distribution of Petal Length. I've colored them by species. This is the base R way. You'll notice that we've had to be clever about how we set the x-axis here, it has to encompass the full range of Petal.Length.

I expect plots to have sensible, nice-looking labels.

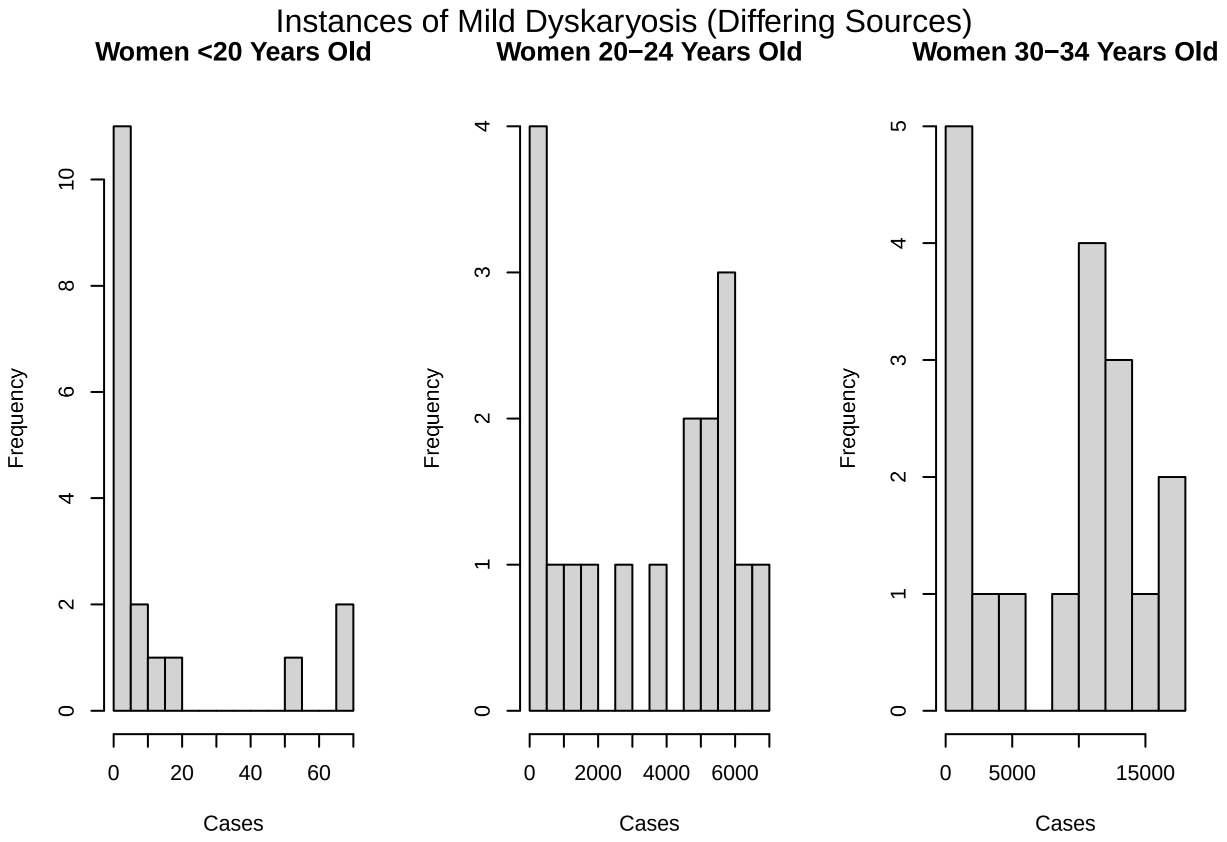
```
#borderline changes for each age group
par(mfrow = c(1, 3))
hist(df1$Borderline_changes, breaks = 10,
     main = "Women <20 Years Old",
     xlab = "Cases")
hist(df2$Borderline_changes, breaks = 10,
     main = "Women 20-24 Years Old",
     xlab = "Cases")
hist(df4$Borderline_changes, breaks = 10,
     main = "Women 30-34 Years Old",
     xlab = "Cases")

mtext("Instances of Borderline Changes (Differing Sources)", outer = TRUE, cex = 1, line = -1.5)
```



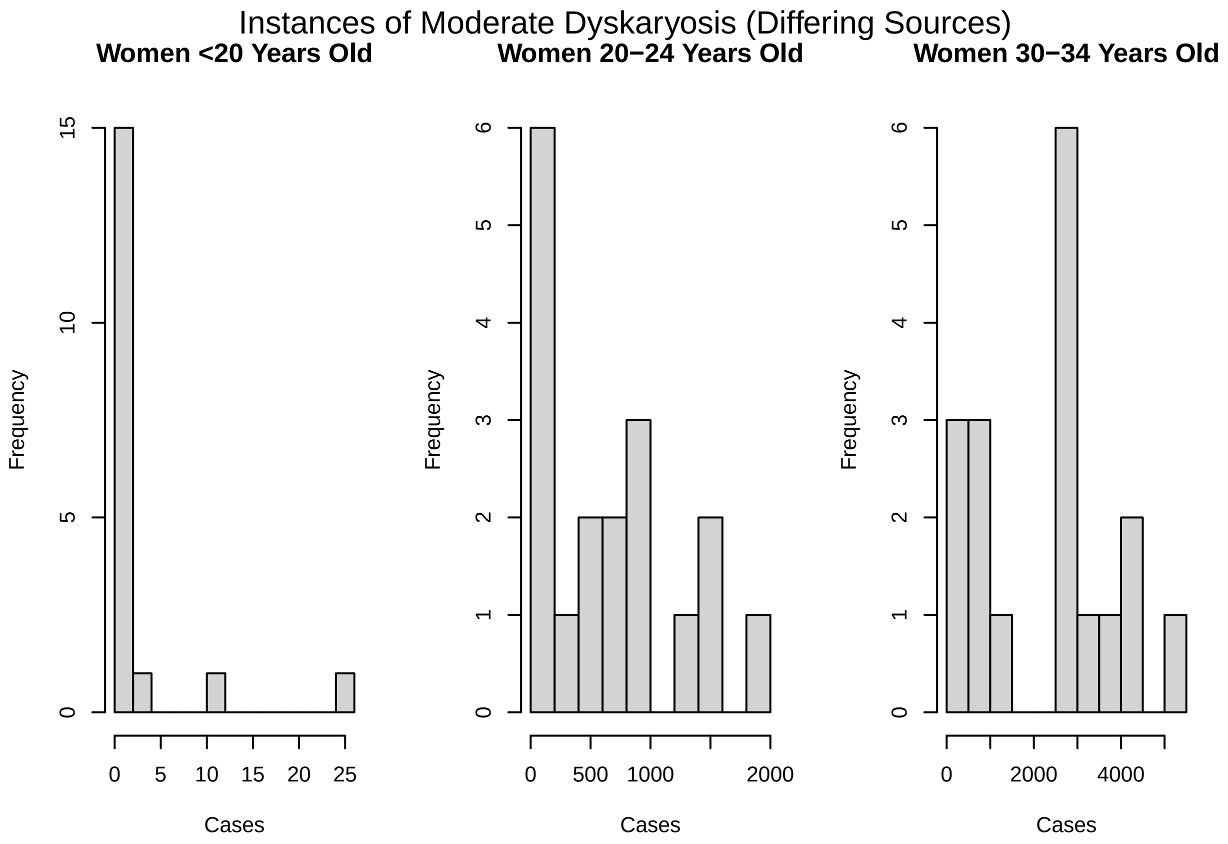
```r
par(mfrow = c(1, 3))
hist(df1$Mild_dyskaryosis, breaks = 10,
     main = "Women <20 Years Old",
     xlab = "Cases")
hist(df2$Mild_dyskaryosis, breaks = 10,
     main = "Women 20-24 Years Old",
     xlab = "Cases")
hist(df4$Mild_dyskaryosis, breaks = 10,
     main = "Women 30-34 Years Old",
     xlab = "Cases")

mtext("Instances of Mild Dyskaryosis (Differing Sources)", outer = TRUE, cex = 1, line = -1.5)
```



```r
par(mfrow = c(1, 3))
hist(df1$Moderate_dyskaryosis, breaks = 10,
     main = "Women <20 Years Old",
     xlab = "Cases")
hist(df2$Moderate_dyskaryosis, breaks = 10,
     main = "Women 20-24 Years Old",
     xlab = "Cases")
hist(df4$Moderate_dyskaryosis, breaks = 10,
     main = "Women 30-34 Years Old",
     xlab = "Cases")

mtext("Instances of Moderate Dyskaryosis (Differing Sources)", outer = TRUE, cex = 1, line = -1.5)
```

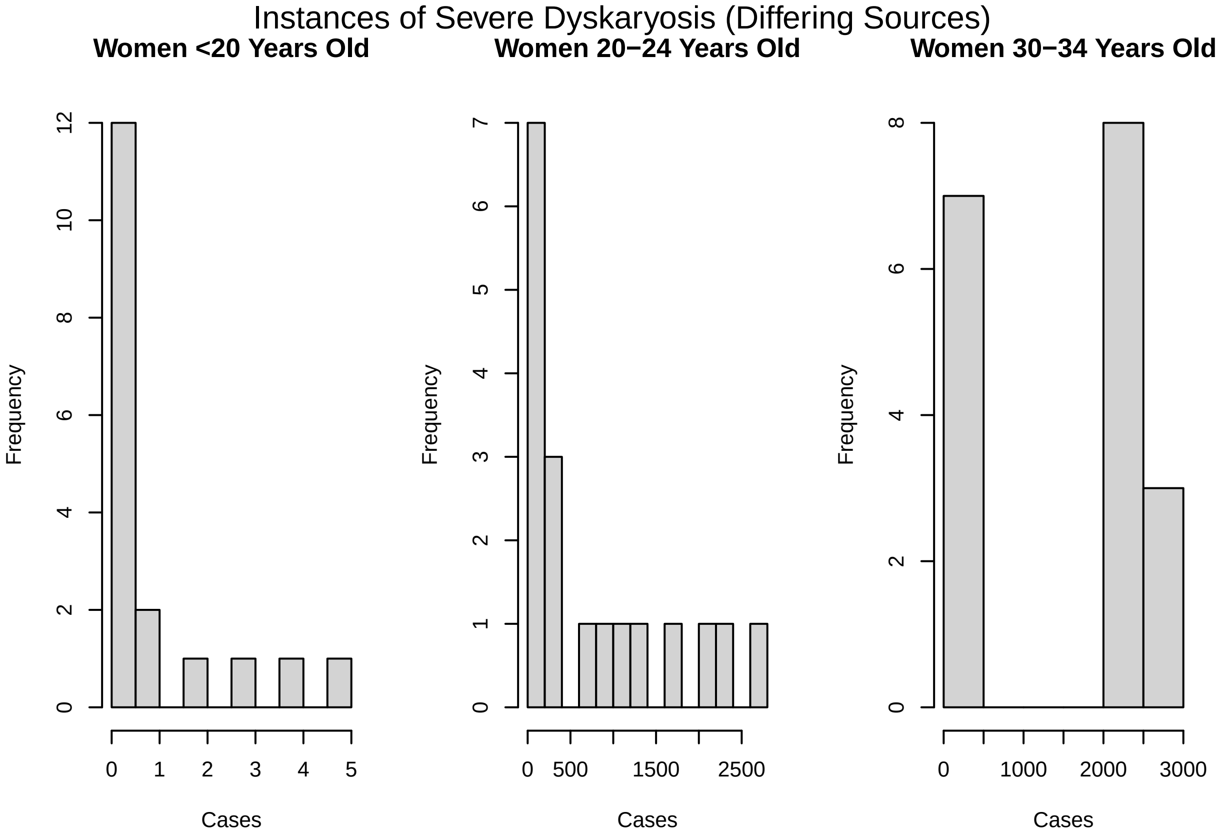


```r
par(mfrow = c(1, 3))
severe_less20 <- severe_NA[severe_NA$Indicator == 1,]
hist(severe_less20$Severe_dyskaryosis, breaks = 10,
     main = "Women <20 Years Old",
     xlab = "Cases")

severe_20_24 <- severe_NA[severe_NA$Indicator == 2,]
hist(severe_20_24$Severe_dyskaryosis, breaks = 10,
     main = "Women 20-24 Years Old",
     xlab = "Cases")

severe_30_34 <- severe_NA[severe_NA$Indicator == 5,]
hist(severe_30_34$Severe_dyskaryosis, breaks = 10,
     main = "Women 30-34 Years Old",
     xlab = "Cases")

mtext("Instances of Severe Dyskaryosis (Differing Sources)", outer = TRUE, cex = 1, line = -1.5)
```



3. Show three scatterplots that show the relationship between variables. Coloring data is a useful dimension to add here! There are many different ways to generate color palettes in R, but the general process here is the same:

- define color palette, here my_colors (or pick your own hex codes or a fancy color brewer package)
- map the variable to the variables. We've done this using fields::color.scale() in class. Or get fancy with for loops and overplotting
- generate plot!

Extremely optional: You can turn your points into emojis with the emojifont package

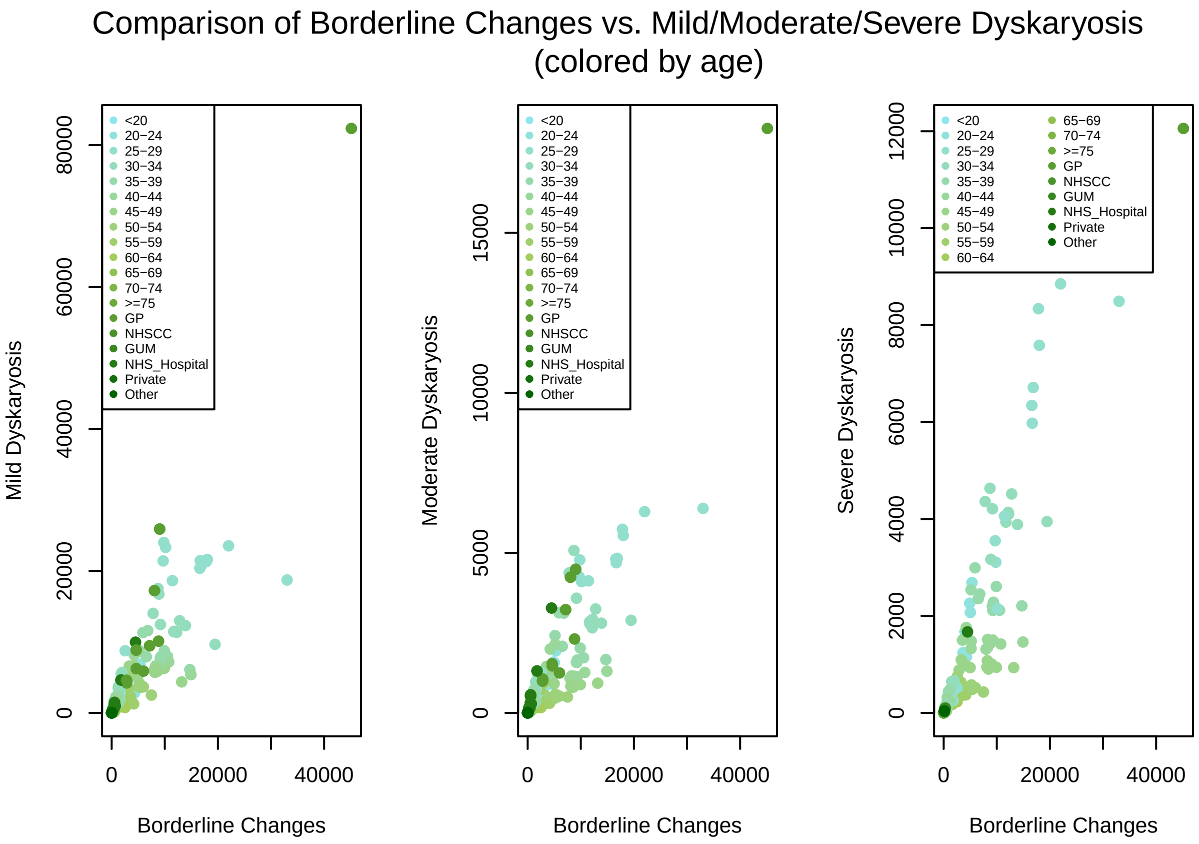
```r
colfunc <- colorRampPalette(c("cadetblue2", "darkolivegreen3", "darkgreen"))
age_color <- fields::color.scale(source$Indicator, col = colfunc(19))

par(mfrow = c(1, 3))
plot(source$Borderline_changes, source$Mild_dyskaryosis, col = age_color,
     pch = 19, xlab = "Borderline Changes", ylab = "Mild Dyskaryosis")
legend("topleft", legend = uniq_ind, pch = 19, col = colfunc(19), cex = .6)

plot(source$Borderline_changes, source$Moderate_dyskaryosis, col = age_color,
     pch = 19, xlab = "Borderline Changes", ylab = "Moderate Dyskaryosis")
legend("topleft", legend = uniq_ind, pch = 19, col = colfunc(19), cex = .6)

plot(source$Borderline_changes, source$Severe_dyskaryosis, col = age_color,
     pch = 19, xlab = "Borderline Changes", ylab = "Severe Dyskaryosis")
legend("topleft", legend = uniq_ind, pch = 19, ncol = 2, col = colfunc(19), cex = .6)
```

```
mtext("Comparison of Borderline Changes vs. Mild/Moderate/Severe Dyskaryosis
      (colored by age)", outer = TRUE, cex = 1, line = -3)
```



4. Write a few sentences about the observations you see in the above plots. Provide any context where necessary.

- We see a general trend of increase in mild to severe dyskaryosis along with borderline changes, showing the increase of more severe cervical cancer cases with the increase of observed borderline changes. The data is colored by age, showing a higher number of cases for younger to mid-age patients. The older patients, along with data collected from differing organizations, are shown to have less cases observed. There is an obvious outlier at the top right corner of each graph, this is because the data collected from GP (general practitioners) have substantially more data compared to other sources and age groups.