Class19

Lucy Wang

Mini-Project: Investigating Pertussis Resurgence

1. Investigating pertussis cases by year

```
# install.packages("datapasta")
# Tools -> Addin -> "Paste as data.frame"
cdc <- data.frame(
                                                  Year = c(1922L,1923L,1
                                                           1925L,1926L,1
                                                           1928L,1929L,
                                                           1930L,1931L,1
                                                           1933L,1934L,1
                                                           1936L,1937L,1
                                                           1939L,1940L,1
                                                           1942L,1943L,1
                                                           1945L,1946L,1
                                                           1948L,1949L,
                                                           1950L,1951L,1
                                                           1953L,1954L,1
                                                           1956L,1957L,1
                                                           1959L,1960L,1
                                                           1962L,1963L,1
                                                           1965L,1966L,1
                                                           1968L,1969L,
                                                           1970L,1971L,1
                                                           1973L,1974L,1
                                                           1976L,1977L,1
                                                           1979L,1980L,1
                                                           1982L,1983L,1
                                                           1985L,1986L,1
```

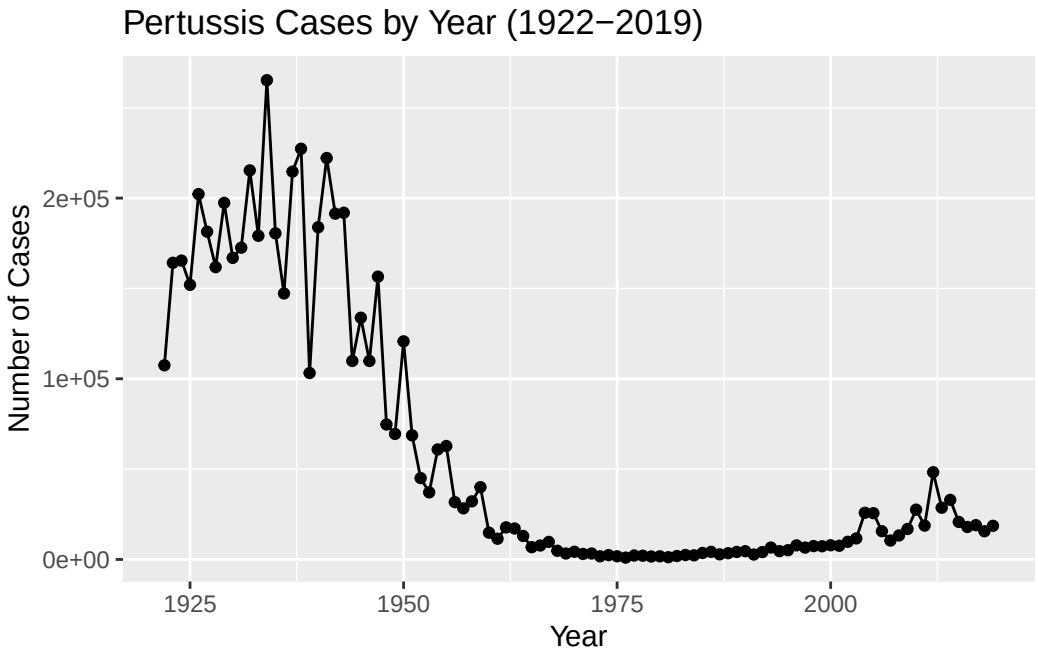
```
                                                    1988L,1989L,
                                                    1990L,1991L,1
                                                    1993L,1994L,1
                                                    1996L,1997L,1
                                                    1999L,2000L,2
                                                    2002L,2003L,2
                                                    2005L,2006L,2
                                                    2008L,2009L,
                                                    2010L,2011L,2
                                                    2013L,2014L,2
                                                    2016L,2017L,2
                                                    2019L),
                                  No..Reported.Pertussis.Cases = c(107473,164191
                                                    165418,152003
                                                    181411,161799
                                                    197371,166914
                                                    215343,179135
                                                    265269,180518
                                                    214652,227319
                                                    103188,183866
                                                    191383,191890
                                                    109873,133792
                                                    156517,74715,
                                                    120718,68687,
                                                    45030,37129,6
                                                    62786,31732,2
                                                    32148,40005,1
                                                    11468,17749,1
                                                    13005,6799,77
                                                    9718,4810,328
                                                    4249,3036,328
                                                    2402,1738,101
                                                    2177,2063,162
                                                    1730,1248,189
                                                    2276,3589,419
                                                    2823,3450,415
                                                    4570,2719,408
                                                    4617,5137,779
                                                    6564,7405,729
                                                    7867,7580,977
                                                    11647,25827,2
```

```
                                                                    15632,10454,1
                                                                    16858,27550,1
                                                                    48277,28639,3
                                                                    20762,17972,1
                                                                    15609,18617)
                                    )
```

Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
library(ggplot2)
p <- ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(title = "Pertussis Cases by Year (1922-2019)", x = "Year", y = "Number of Cases")

p
```

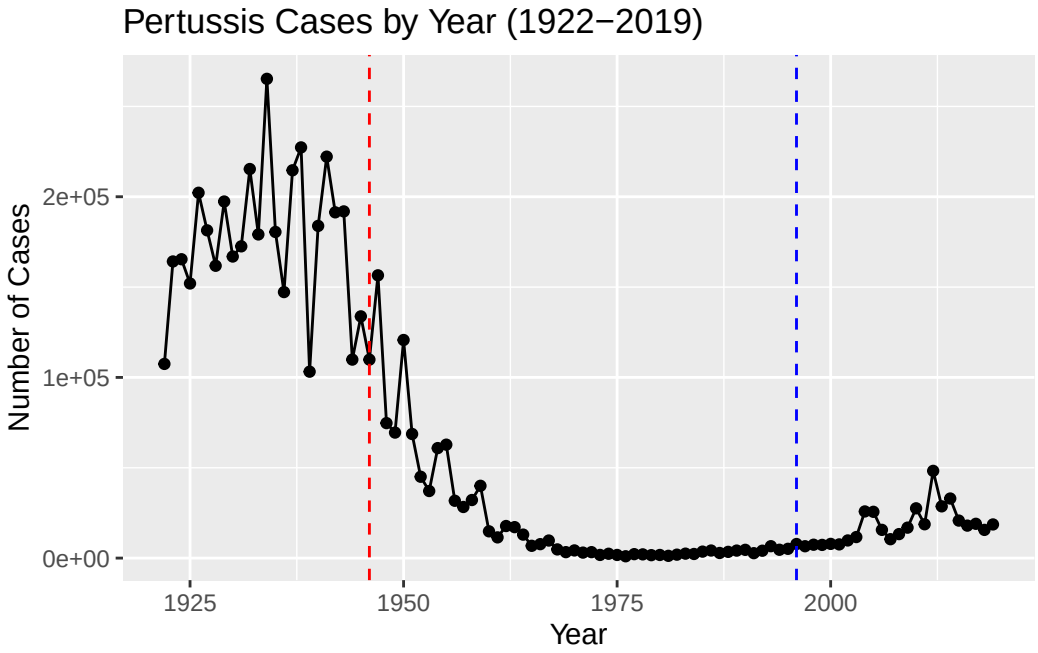


2. A tale of two vaccines (wP & aP)

Examine what happened after the switch to the acellular pertussis (aP) vaccination program, adding lines to plot.

> Q2. Using the ggplot geom_vline() function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice?

```
# Adding geom_vline (xintercept = 1946 & 1996)
p +
  geom_vline(xintercept = 1946, color = "red", linetype = 2) +
  geom_vline(xintercept = 1996, color = "blue", linetype = 2)
```



> Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend?

There is a raise in number of cases after the introduction of aP vaccine. It is possible that due to more frequent testing, the disease is diagnosed more frequently. Maybe the pathogen evolved to escape some of human immune mechanisms from decades of attempts to invade human.

3. Exploring CMI-PB data

```
# Allows us to read, write and process JSON data
library(jsonlite)
```

```
# Read table from CMI-PB API
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
```

Q4. How may aP and wP infancy vaccinated subjects are in the dataset?

```
table(subject$infancy_vac)
```

```
aP wP
47 49
```

Q5. How many Male and Female subjects/patients are in the dataset?

```
table(subject$biological_sex)
```

```
Female   Male
    66     30
```

Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc...)?

```
table(subject$biological_sex, subject$race)
```

```
        American Indian/Alaska Native Asian Black or African American
  Female                             0    18                        2
  Male                               1     9                        0

        More Than One Race Native Hawaiian or Other Pacific Islander
  Female                  8                                        1
  Male                    2                                        1

        Unknown or Not Reported White
  Female                      10    27
  Male                         4    13
```

Side-Note: Working with dates

```
# There is a package written for dates operations
library(lubridate)
```

```
Attaching package: 'lubridate'
```

```
The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2023-03-16"
```

```
# To Calculate time differences
today() - ymd("2000-01-01")  # In days
```

```
Time difference of 8475 days
```

```
time_length( today() - ymd("2000-01-01"),  "years") # In years
```

```
[1] 23.20329
```

Q7. Using this approach determine (i) the average age of wP individuals, (ii) the average age of aP individuals; and (iii) are they significantly different?

```
# Use todays date to calculate age in days
subject$age <- today() - ymd(subject$year_of_birth)
time_length(subject$age, "years")
```

```
 [1] 37.20192 55.20329 40.20260 35.20329 32.20260 35.20329 42.20123 38.20123
 [9] 27.20329 41.20192 37.20192 41.20192 26.20123 30.20123 34.20123 36.20260
[17] 43.20329 26.20123 29.20192 36.20260 30.20123 28.20260 30.20123 33.20192
[25] 47.20329 51.20329 51.20329 33.20192 25.20192 25.20192 32.20260 28.20260
```

```
[33] 28.20260 25.20192 25.20192 35.20329 30.20123 36.20260 31.20329 30.20123
[41] 25.20192 24.20260 26.20123 23.20329 25.20192 23.20329 23.20329 26.20123
[49] 24.20260 25.20192 23.20329 27.20329 24.20260 25.20192 23.20329 42.20123
[57] 40.20260 38.20123 32.20260 31.20329 35.20329 40.20260 26.20123 41.20192
[65] 26.20123 35.20329 34.20123 26.20123 33.20192 40.20260 32.20260 26.20123
[73] 25.20192 26.20123 38.20123 29.20192 38.20123 26.20123 25.20192 25.20192
[81] 26.20123 25.20192 27.20329 25.20192 26.20123 26.20123 26.20123 25.20192
[89] 25.20192 26.20123 26.20123 26.20123 27.20329 26.20123 26.20123 26.20123
```

```r
# aP average ages
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
ap <- subject %>% filter(infancy_vac == "aP")

round( summary( time_length( ap$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     23      25      26      26      26      27
```

```r
# wP average ages
wp <- subject %>% filter(infancy_vac == "wP")
round( summary( time_length( wp$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     28      32      35      36      40      55
```

They are significantly different.

Q8. Determine the age of all individuals at time of boost?

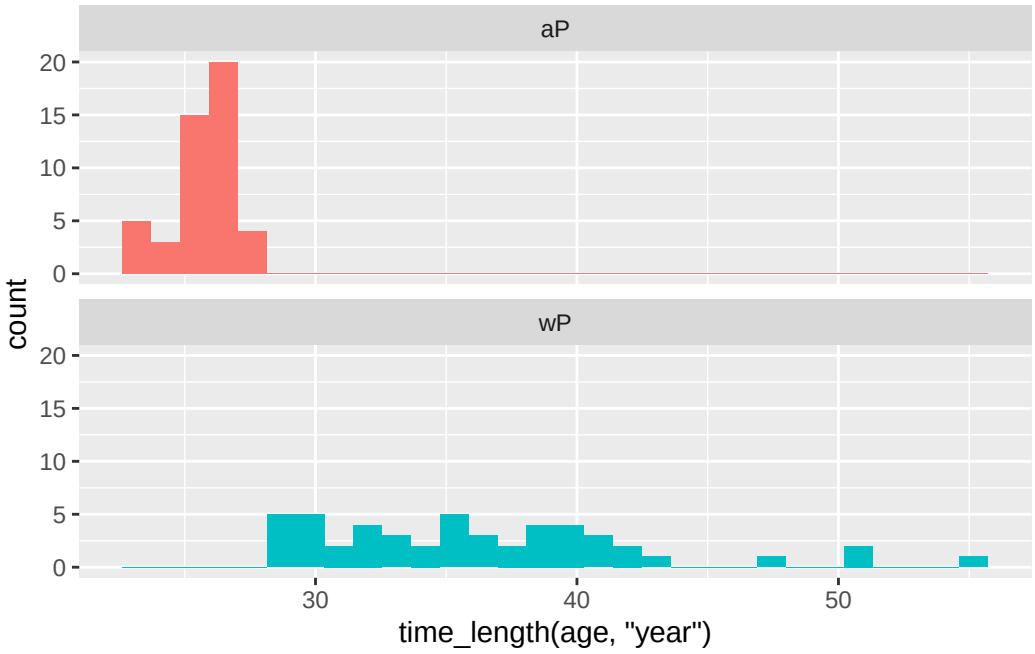
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

Q9. With the help of a faceted boxplot (see below), do you think these two groups are significantly different?

```
ggplot(subject) +
  aes(time_length(age, "year"),
      fill=as.factor(infancy_vac)) +
  geom_histogram(show.legend=FALSE) +
  facet_wrap(vars(infancy_vac), nrow=2)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Yes, I think it is significantly different since there is little overlap between two set of data, which gives very different mean and median.

Joining multiple tables

```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/ab_titer", simplifyVector = TRUE)
```

Q9. Complete the code to join specimen and subject tables to make a new merged data frame containing all specimen records along with their associated subject details:

```
meta <- full_join(specimen, subject)
```

```
Joining with `by = join_by(subject_id)`
```

```
dim(meta)
```

```
[1] 729  14
```

```
head(meta)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                          736
3           3          1                            1
4           4          1                            3
5           5          1                            7
6           6          1                           11
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                           736         Blood    10          wP         Female
3                             1         Blood     2          wP         Female
4                             3         Blood     3          wP         Female
5                             7         Blood     4          wP         Female
6                            14         Blood     5          wP         Female
               ethnicity  race year_of_birth date_of_boost      dataset
1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
```

```
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
         age
1 13588 days
2 13588 days
3 13588 days
4 13588 days
5 13588 days
6 13588 days
```

Q10. Now using the same procedure join meta with titer data so we can further analyze this data in terms of time of visit aP/wP, male/female etc.

```
abdata <- inner_join(titer, meta, by = "specimen_id")
dim(abdata)
```

```
[1] 32675    21
```

Q11. How many specimens (i.e. entries in abdata) do we have for each isotype?

```
# Use table() on isotype column
table(abdata$isotype)
```

```
 IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 1413 6141 6141 6141 6141
```

Q12. What do you notice about the number of visit 8 specimens compared to other visits?

```
table(abdata$visit)
```

```
   1    2    3    4    5    6    7    8
5795 4640 4640 4640 4640 4320 3920   80
```

8 specimens are far less than the other visits.

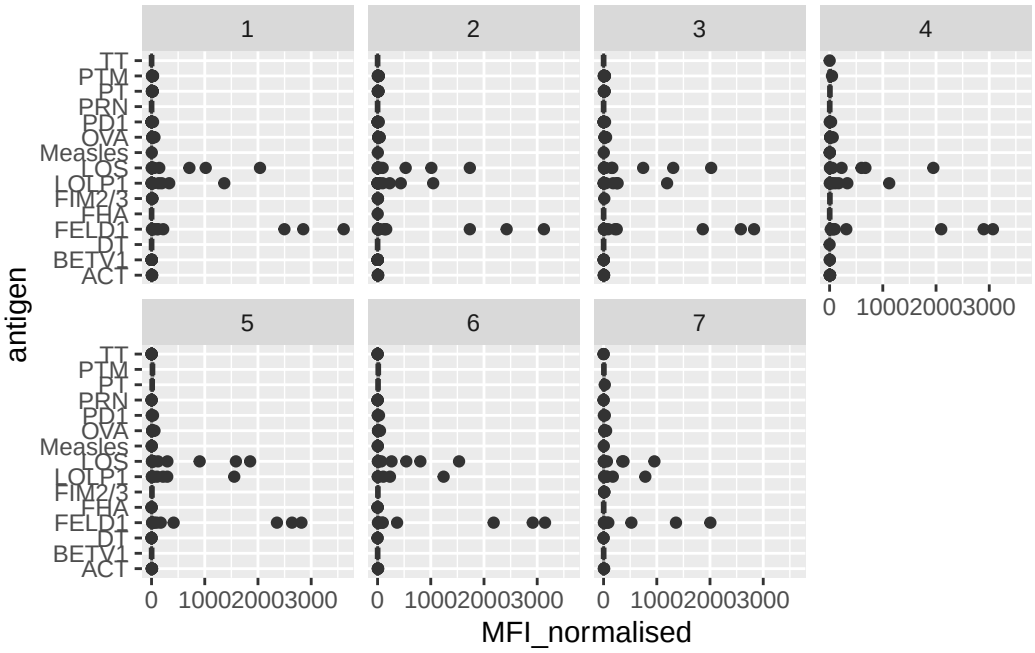
4. Examine IgG1 Ab titer levels

```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1    IgG1                TRUE     ACT 274.355068      0.6928058
2           1    IgG1                TRUE     LOS  10.974026      2.1645083
3           1    IgG1                TRUE   FELD1   1.448796      0.8080941
4           1    IgG1                TRUE   BETV1   0.100000      1.0000000
5           1    IgG1                TRUE   LOLP1   0.100000      1.0000000
6           1    IgG1                TRUE Measles  36.277417      1.6638332
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 IU/ML                 3.848750          1                           -3
2 IU/ML                 4.357917          1                           -3
3 IU/ML                 2.699944          1                           -3
4 IU/ML                 1.734784          1                           -3
5 IU/ML                 2.550606          1                           -3
6 IU/ML                 4.438966          1                           -3
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                             0         Blood     1          wP         Female
3                             0         Blood     1          wP         Female
4                             0         Blood     1          wP         Female
5                             0         Blood     1          wP         Female
6                             0         Blood     1          wP         Female
              ethnicity  race year_of_birth date_of_boost      dataset
1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
        age
1 13588 days
2 13588 days
3 13588 days
4 13588 days
5 13588 days
6 13588 days
```

Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

```
ggplot(ig1) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

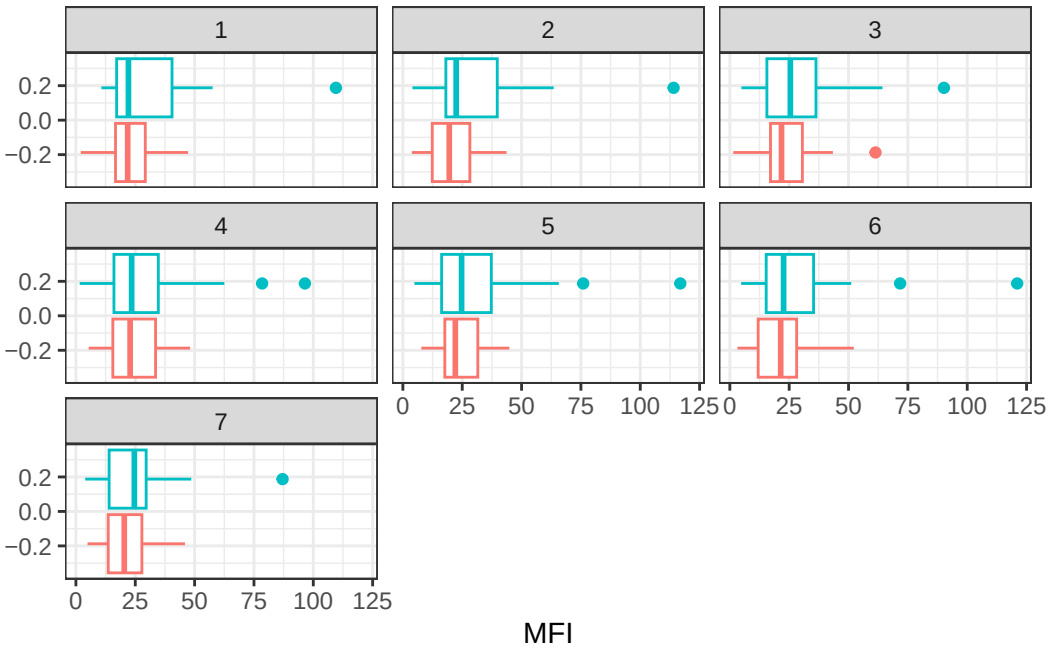
Mainly FIM2/3, also FHA and DT show slight increase. Maybe these are epitopes of pathogenic characteristic, so IgG1 could recognize them overtime and become increasingly accurate.

We can attempt to examine differences between wP and aP here by setting color and/or facet values of the plot to include `infancy_vac` status. However these plots tend to be rather busy and thus hard to interpret easily.

Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate in substrate attachment).

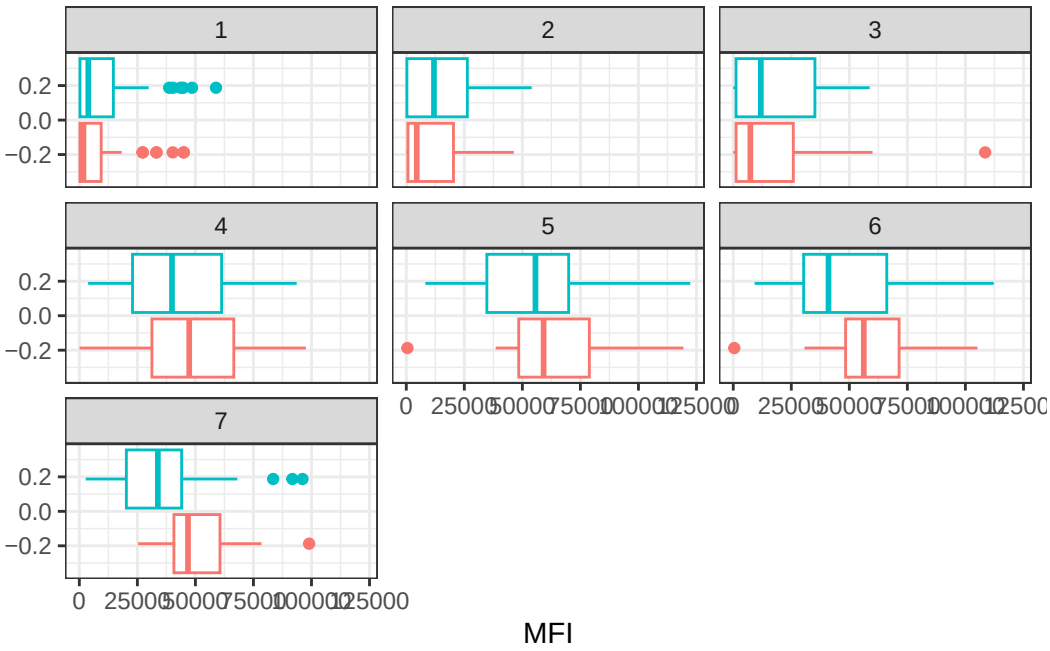
```
# For measles
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = "Measles") +
  facet_wrap(vars(visit)) +
  theme_bw()
```

```
Warning: `show.legend` must be a logical vector.
```




```
# For FIM2/3
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = "FIM2/3") +
  facet_wrap(vars(visit)) +
  theme_bw()
```

```
Warning: `show.legend` must be a logical vector.
```



Q16. What do you notice about these two antigens time course and the FIM2/3 data in particular?

FIM2/3 rapidly increases from visit 1 and is far more than measles. It reaches its maximum value in visit 5.

Q17. Do you see any clear difference in aP vs. wP responses?

wP seems to show less a trend of increase and stays lower when aP increases. Maybe indicating that aP is more responsive during the visits.

5. Obtaining CMI-PB RNASeq data

Investigating IgG1 gene

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.
```

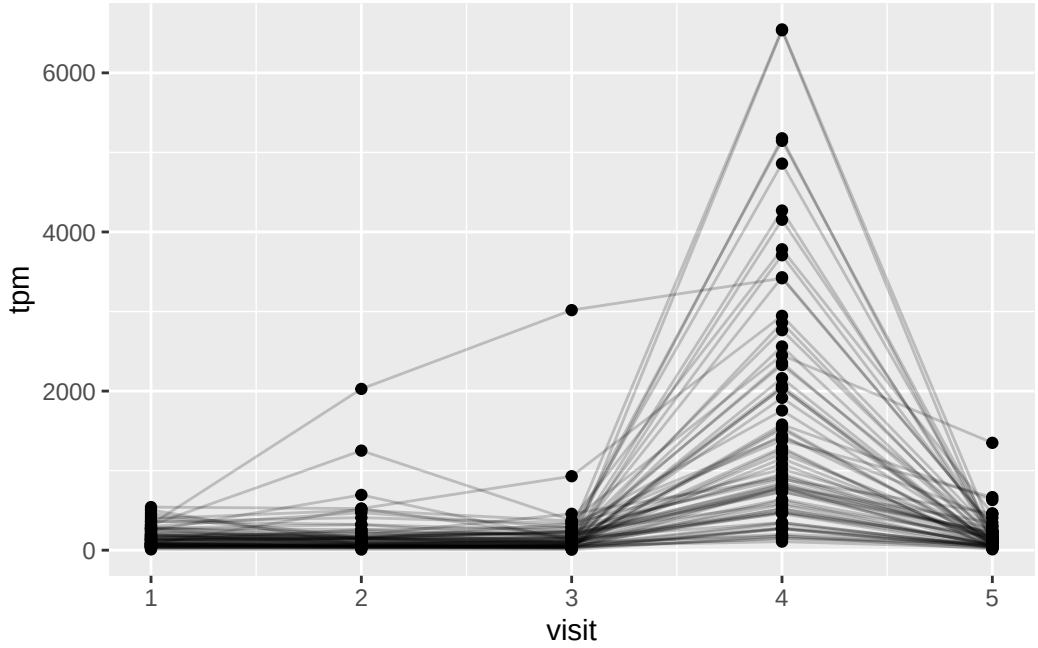
```
rna <- read_json(url, simplifyVector = TRUE)
```

```
#Just like meta <- inner_join(specimen, subject)
ssrna <- inner_join(rna, meta)
```

```
Joining with `by = join_by(specimen_id)`
```

Q18. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of `visit` vs. `tpm`).

```
ggplot(ssrna) +
  aes(visit, tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```



Q19.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

It peaks at visit 4.

Q20. Does this pattern in time match the trend of antibody titer data? If not, why not?

I think it matches the antibody titer, though the antibody peaks at visit 5 instead of visit 4. We can regard the production of antibody a response to the expression of those genes, which can take 1 visit.

-The End -