Pertussis and the CMI-PB project

Benjie Miao (A69026849)

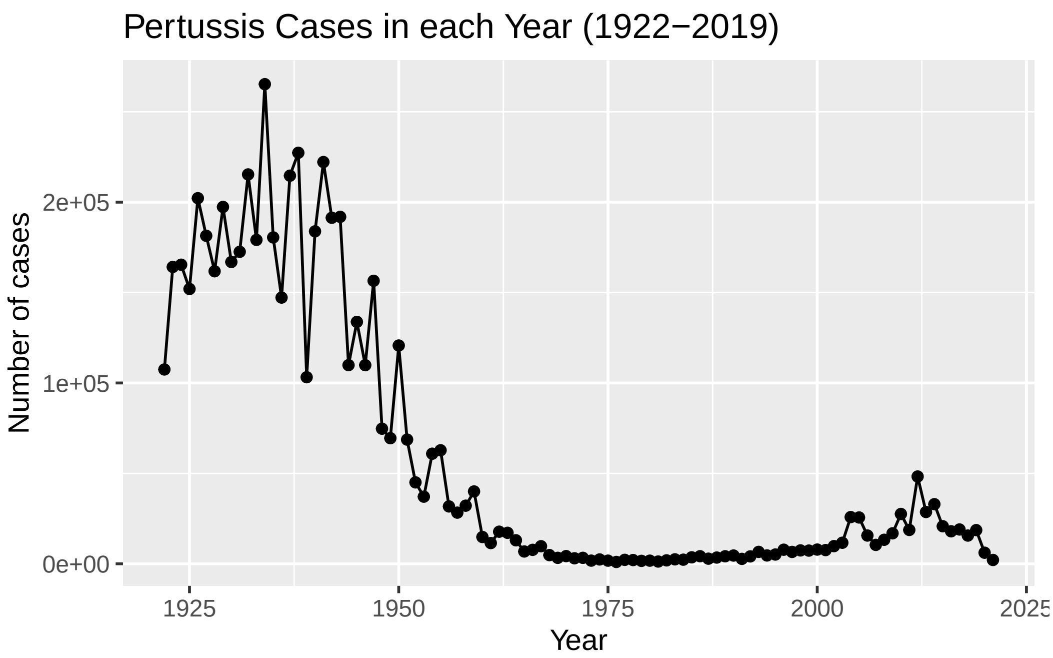
1. Investigating pertussis cases by year

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(datapasta)
library(ggplot2)

cases <- read.csv("case_by_year.csv")
head(cases)
```

```
  Year No..Reported.Pertussis.Cases
1 1922                       107473
2 1923                       164191
3 1924                       165418
4 1925                       152003
5 1926                       202210
6 1927                       181411
```

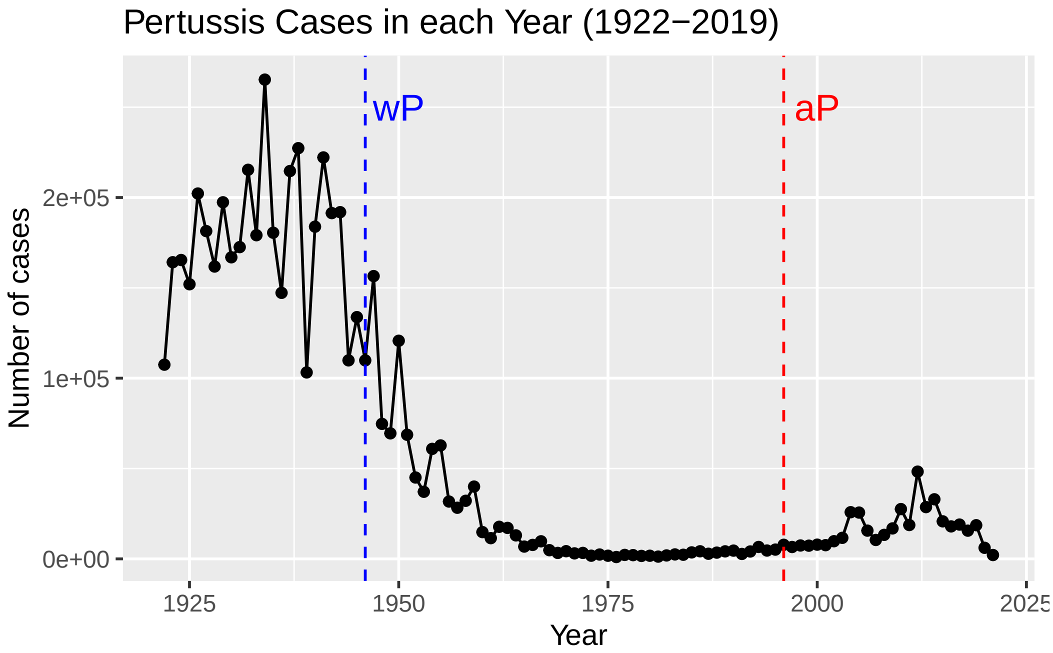
```
plot <- ggplot(cases, aes(Year, No..Reported.Pertussis.Cases)) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y = "Number of cases",
  title = "Pertussis Cases in each Year (1922-2019)")
plot
```



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

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?

```
plot_vline <- plot +
  geom_vline(xintercept = 1946, linetype = "dashed", col = "blue") +
  geom_vline(xintercept = 1996, linetype = "dashed", col = "red") +
  annotate(geom = "text", x = 1950, y = 250000, label = "wP", col = "blue", size = 5) +
  annotate(geom = "text", x = 2000, y = 250000, label = "aP", col = "red", size = 5)
plot_vline
```



The wP application dramatically decreased the cases, while the ap slightly increase the number of cases.

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

The case number increases. It may be due to that the effect of aP vaccine is not effective as wP vaccine.

3. Exploring CMI-PB data

```
# Allows us to read, write and process JSON data
library(jsonlite)
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
head(subject, 3)
```

```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          2          wP         Female Not Hispanic or Latino White
3          3          wP         Female                Unknown White
```

```
  year_of_birth date_of_boost      dataset
1    1986-01-01    2016-09-12 2020_dataset
2    1968-01-01    2019-01-28 2020_dataset
3    1983-01-01    2016-10-10 2020_dataset
```

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

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

```

aP wP 
60 58 
```

60 and 58.

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

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

```

Female   Male 
    79     39 
```

79 and 39.

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    21                         2
  Male                               1    11                         0
        
         More Than One Race Native Hawaiian or Other Pacific Islander
  Female                  9                                         1
  Male                    2                                         1
```

```
         Unknown or Not Reported White
  Female                      11    35
  Male                         4    20
```

Side-Note: Working with dates

```r
library(lubridate)
```

```
Attaching package: 'lubridate'


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

    date, intersect, setdiff, union
```

```r
today()
```

```
[1] "2023-12-11"
```

```r
today() - ymd("2000-01-01")
```

```
Time difference of 8745 days
```

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

```
[1] 23.94251
```

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?

```r
subject$age <- time_length(today() - ymd(subject$year_of_birth), "years")

age_wp <- subject[subject$infancy_vac == "wP", "age"]
age_ap <- subject[subject$infancy_vac == "aP", "age"]
```

```
mean(age_wp)
```

```
[1] 36.33798
```

```
mean(age_ap)
```

```
[1] 26.04125
```

```
t.test(age_wp, age_ap)
```

```

	Welch Two Sample t-test

data:  age_wp and age_ap
t = 12.436, df = 65.411, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.643385 11.950080
sample estimates:
mean of x mean of y 
 36.33798  26.04125 
```

36. 26. Yes they are significantly different.

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

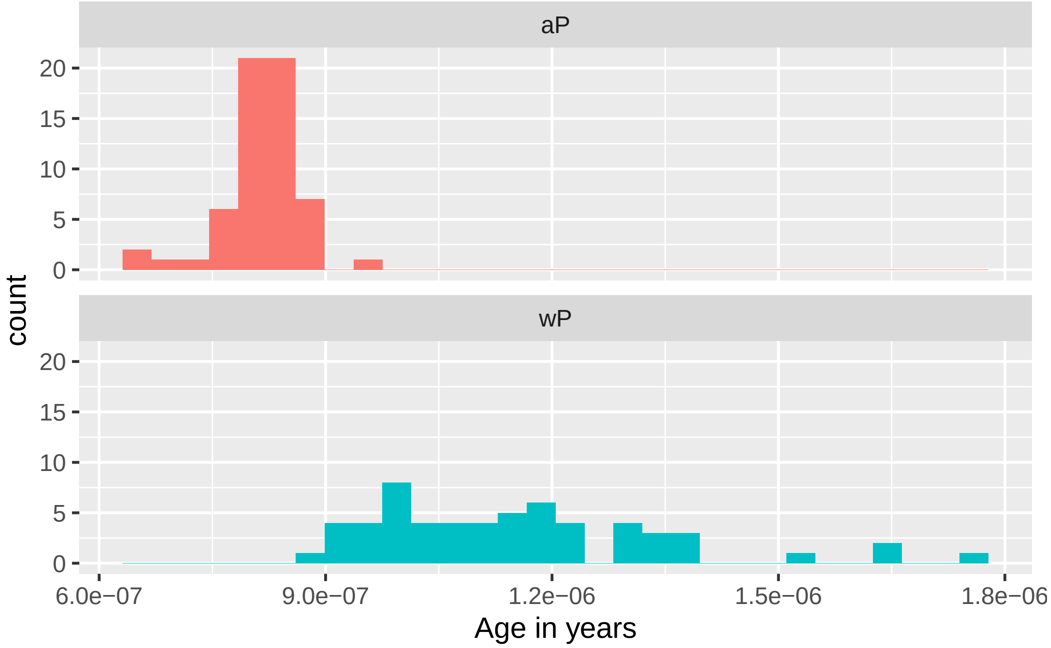
```
age_at_boost  <- time_length(ymd(subject$date_of_boost) - ymd(subject$year_of_birth), "yea
head(age_at_boost)
```

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

Q9. With the help of a faceted boxplot or histogram (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) +
  xlab("Age in years")
```

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



Joining multiple tables

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

```
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
```

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

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

```
dim(meta)
```

```
[1] 939  14
```

```
head(meta)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                            1
3           3          1                            3
4           4          1                            7
5           5          1                           11
6           6          1                           32
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                             1         Blood     2          wP         Female
3                             3         Blood     3          wP         Female
4                             7         Blood     4          wP         Female
5                            14         Blood     5          wP         Female
6                            30         Blood     6          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 37.94114
2 37.94114
3 37.94114
4 37.94114
5 37.94114
6 37.94114
```

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)
```

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

```
dim(abdata)
```

```
[1] 41810    21
```

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

```
table(abdata$isotype)
```

```

 IgE  IgG IgG1 IgG2 IgG3 IgG4 
6698 3240 7968 7968 7968 7968 
```

Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most “recent” dataset?

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

```

   1    2    3    4    5    6    7    8 
6390 6460 6530 5900 5900 5475 5075   80 
```

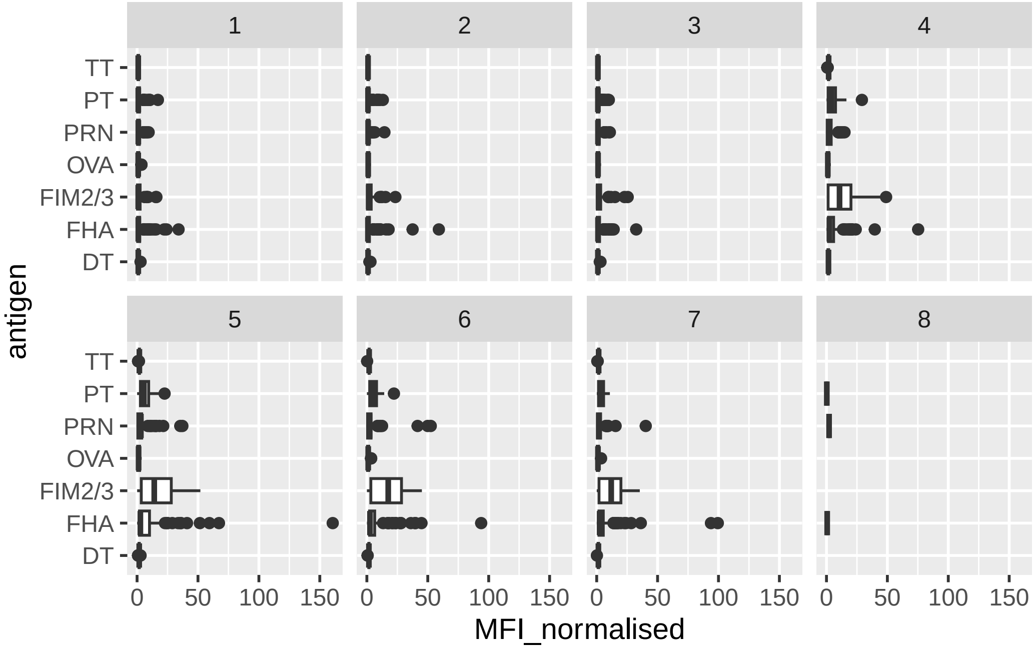
4. Examine IgG Ab titer levels

```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
4          19     IgG                TRUE      PT   20.11607       1.096366
5          19     IgG                TRUE     PRN  976.67419       7.652635
6          19     IgG                TRUE     FHA   60.76626       1.096457
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 IU/ML                 0.530000          1                           -3
2 IU/ML                 6.205949          1                           -3
3 IU/ML                 4.679535          1                           -3
4 IU/ML                 0.530000          3                           -3
5 IU/ML                 6.205949          3                           -3
6 IU/ML                 4.679535          3                           -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                Unknown White    1983-01-01    2016-10-10 2020_dataset
5                Unknown White    1983-01-01    2016-10-10 2020_dataset
6                Unknown White    1983-01-01    2016-10-10 2020_dataset
       age
1 37.94114
2 37.94114
3 37.94114
4 40.94182
5 40.94182
6 40.94182
```

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

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

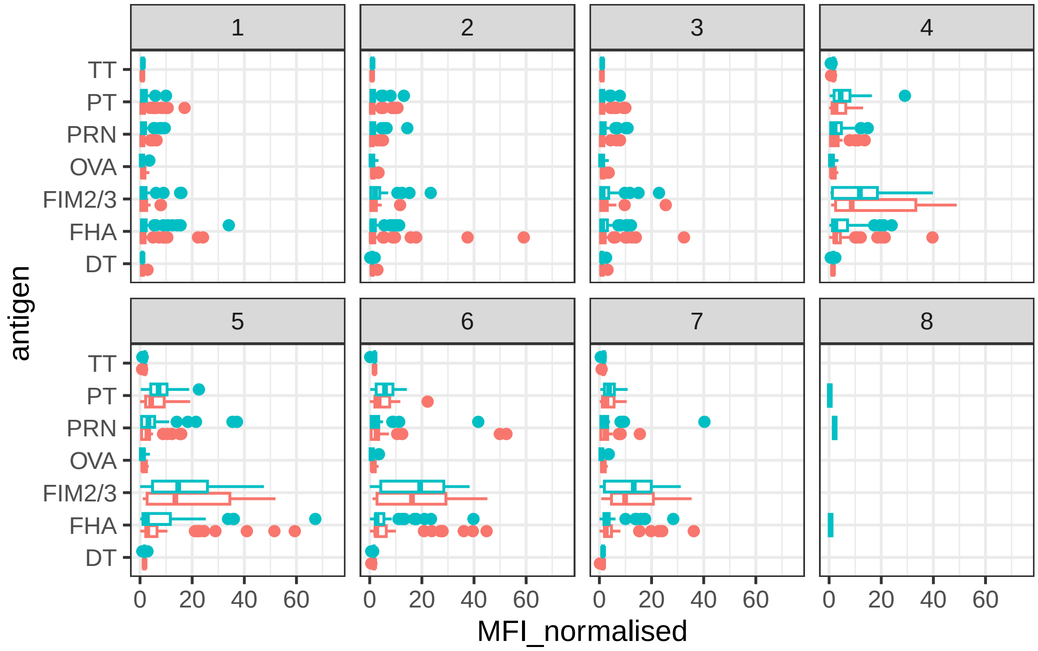


Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

FIM 2/3 shows differences along the longtitudial axis.

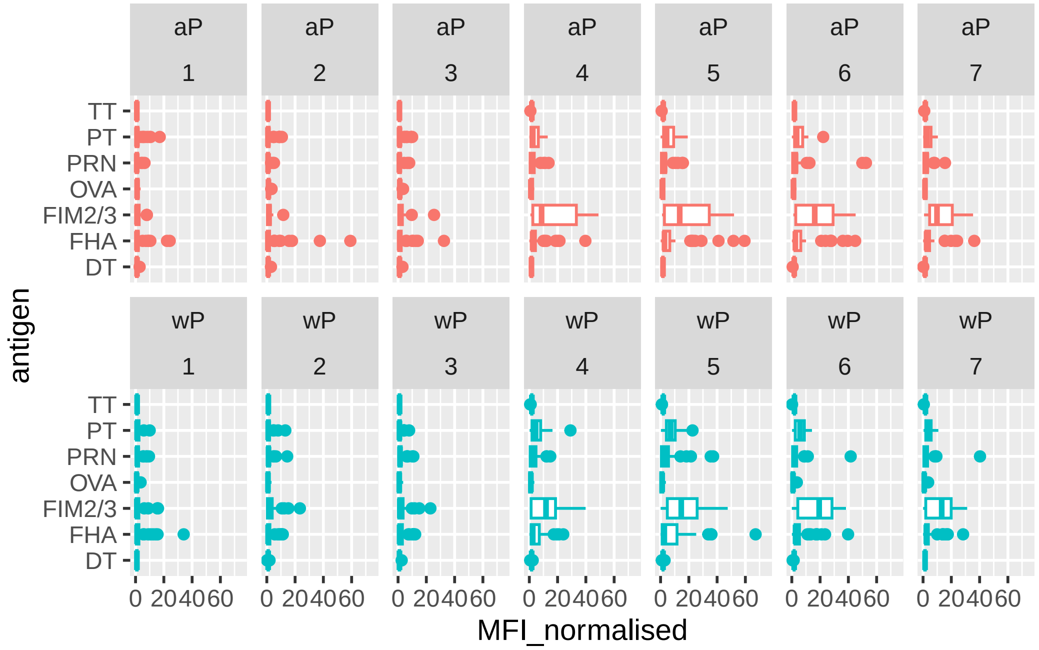
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

```
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).
```



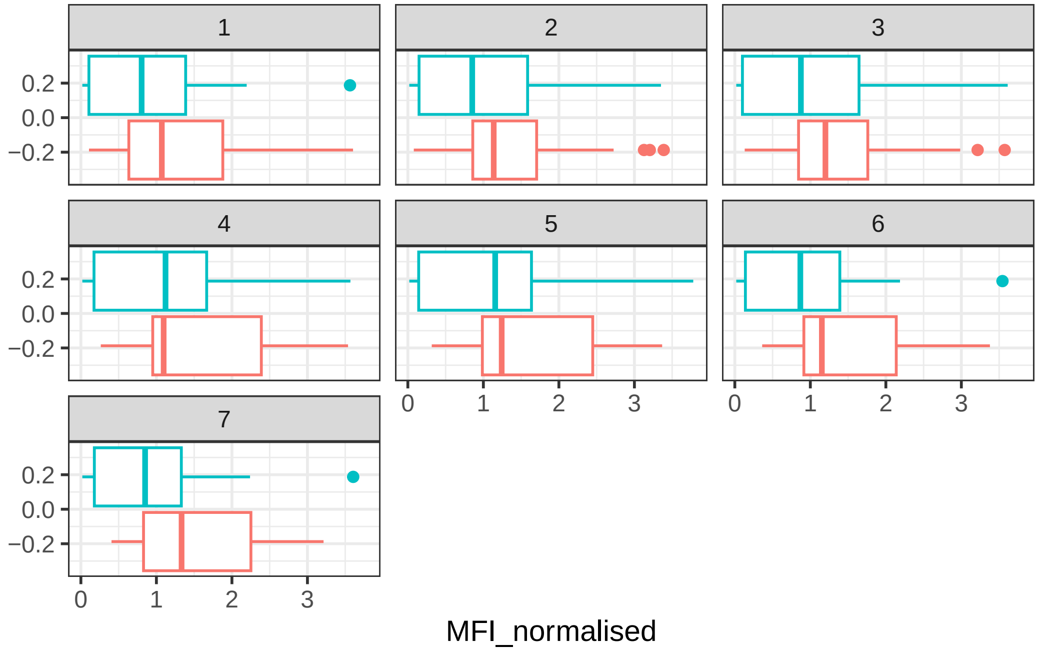
```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

```
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).
```

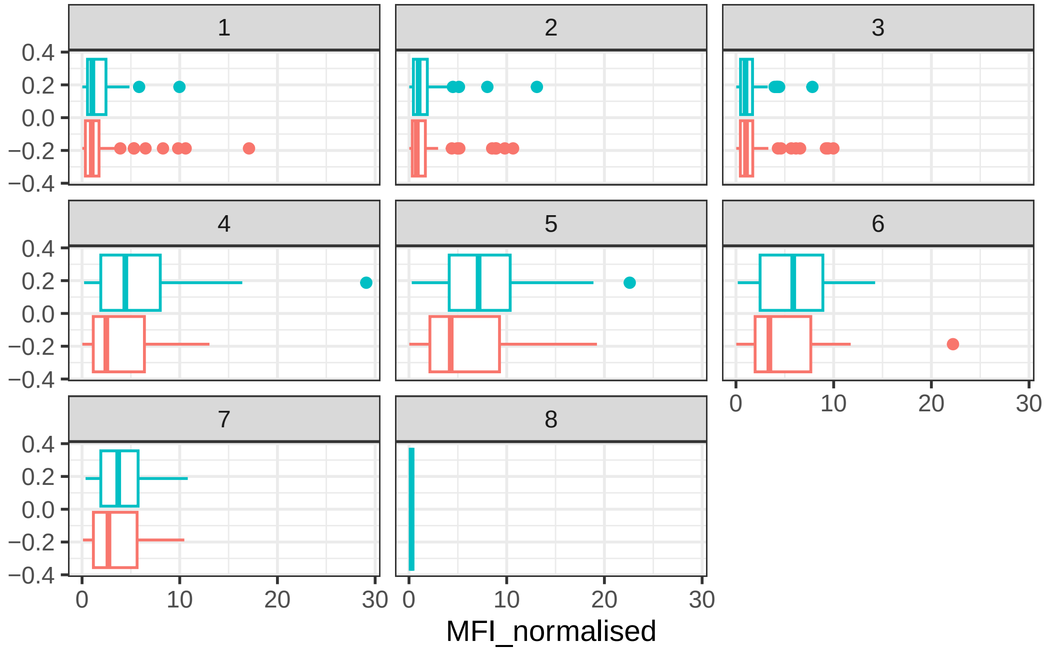


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium B. pertussis).

```
filter(igg, antigen=="OVA") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



```
filter(igg, antigen=="PT") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Q16. What do you notice about these two antigens time courses and the PT data in particular?

The PT level reaches the peak at the 5th visit. This trend is consistent across wP/aP.

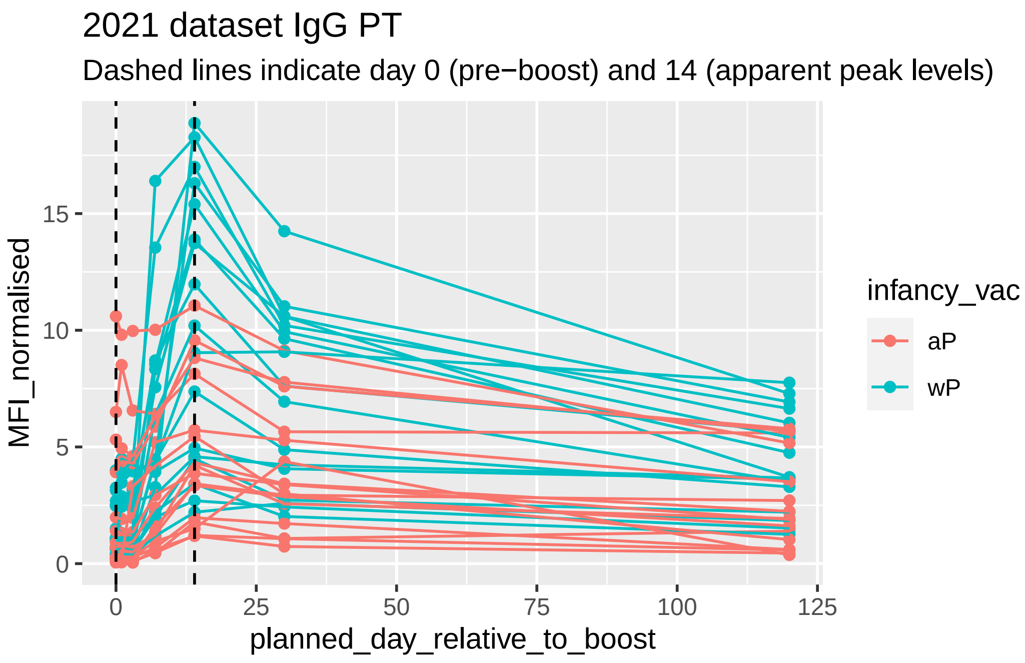
Q17. Do you see any clear difference in aP vs. wP responses?

```
t.test(filter(igg, antigen == "FIM2/3" & infancy_vac == "wP")$MFI,
       filter(igg, antigen == "FIM2/3" & infancy_vac == "aP")$MFI)
```

```

    Welch Two Sample t-test

data:  filter(igg, antigen == "FIM2/3" & infancy_vac == "wP")$MFI and filter(igg, antigen ==
t = -0.26979, df = 280.24, p-value = 0.7875
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1717.676  1303.597
sample estimates:
mean of x mean of y
 4495.841  4702.881
```

Though wP is higher than aP, the value has no significant difference.

```
abdata.21 <- abdata %>% filter(dataset == "2021_dataset")

abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT") %>%
  ggplot() +
    aes(x=planned_day_relative_to_boost,
        y=MFI_normalised,
        col=infancy_vac,
        group=subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
  labs(title="2021 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```



Q18. Does this trend look similar for the 2020 dataset?

Yes.

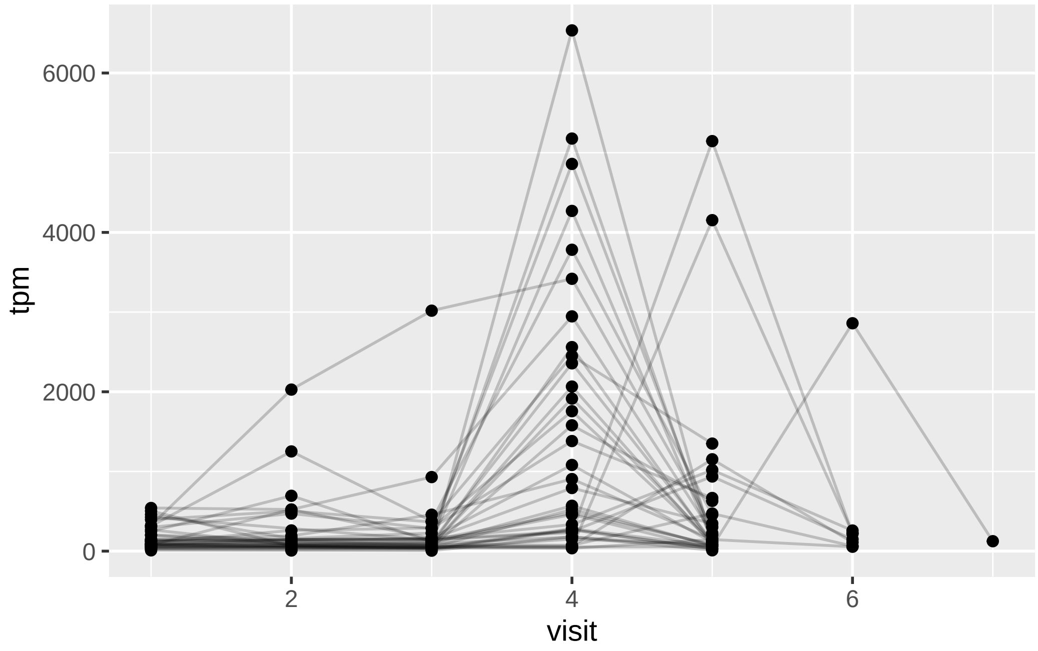
5. Obtaining CMI-PB RNASeq data

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.
rna <- read_json(url, simplifyVector = TRUE)
#meta <- inner_join(specimen, subject)
ssrna <- inner_join(rna, meta)
```

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

Q19. 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)
```



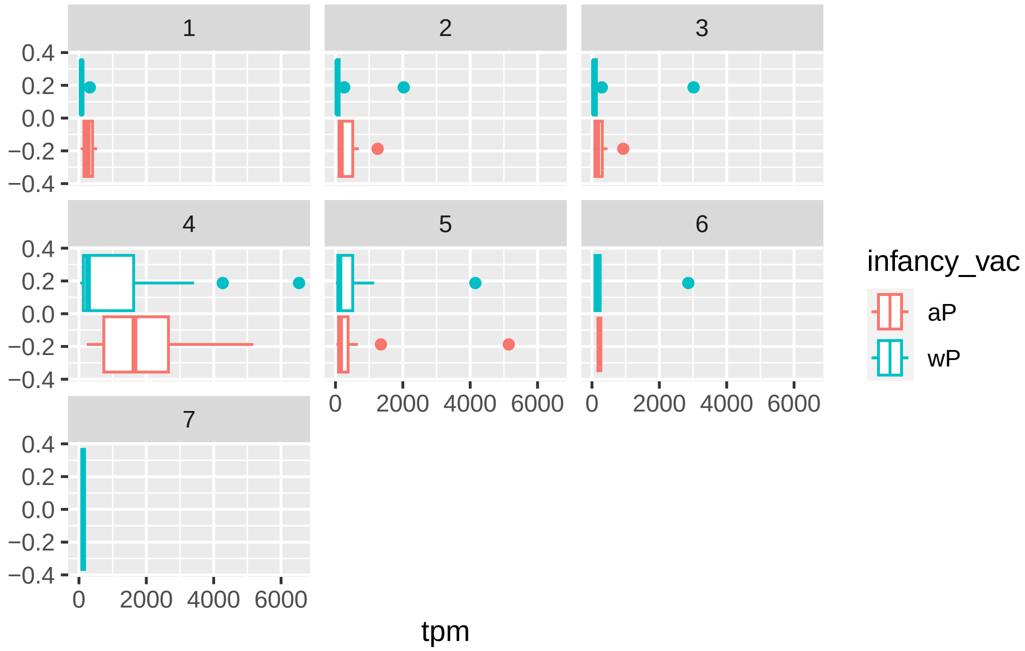
Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

The expression reaches the peak at the 4th visit.

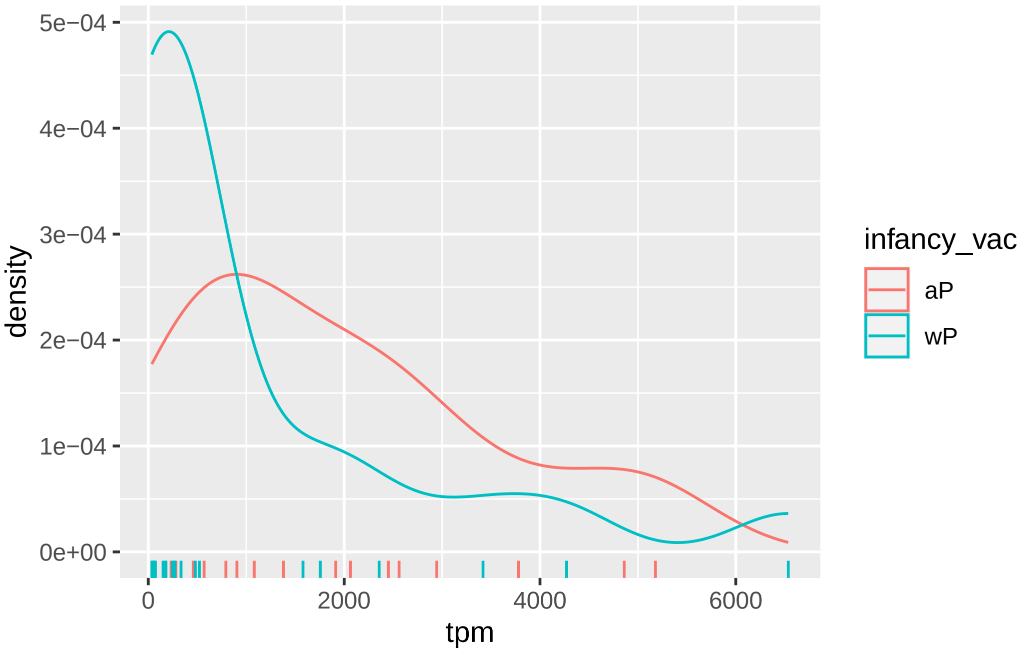
Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

Slightly differenct since the antibody reaches the peak at 5th. It may be due to that the antibody has a lagging effect than the gene expression.

```
ggplot(ssrna) +
  aes(tpm, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(visit))
```



```
ssrna %>%
  filter(visit==4) %>%
  ggplot() +
    aes(tpm, col=infancy_vac) + geom_density() +
    geom_rug()
```


5e−04
4e−04
3e−04
2e−04
1e−04
0e+00
density
0
2000
4000
6000
tpm
infancy_vac
aP
wP