Class18

Mason Lew (PID:A17533139)

2025-03-12

Table of contents

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.

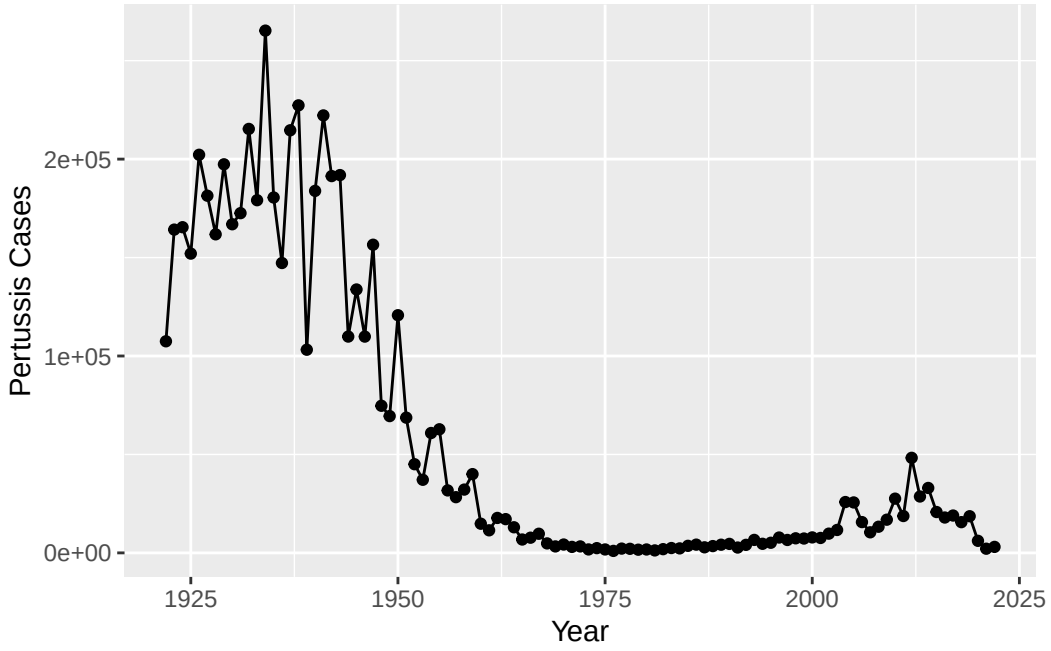
```
cdc <- data.frame(
                            Year = c(1922L,
                                     1923L,1924L,1925L,1926L,1927L,1928L,
                                     1929L,1930L,1931L,1932L,1933L,1934L,1935L,
                                     1936L,1937L,1938L,1939L,1940L,1941L,
                                     1942L,1943L,1944L,1945L,1946L,1947L,1948L,
                                     1949L,1950L,1951L,1952L,1953L,1954L,
                                     1955L,1956L,1957L,1958L,1959L,1960L,
                                     1961L,1962L,1963L,1964L,1965L,1966L,1967L,
                                     1968L,1969L,1970L,1971L,1972L,1973L,
                                     1974L,1975L,1976L,1977L,1978L,1979L,1980L,
                                     1981L,1982L,1983L,1984L,1985L,1986L,
                                     1987L,1988L,1989L,1990L,1991L,1992L,1993L,
```

```
                                    1994L,1995L,1996L,1997L,1998L,1999L,
                                    2000L,2001L,2002L,2003L,2004L,2005L,
                                    2006L,2007L,2008L,2009L,2010L,2011L,2012L,
                                    2013L,2014L,2015L,2016L,2017L,2018L,
                                    2019L,2020L,2021L,2022L),
  No..Reported.Pertussis.Cases = c(107473,
                                    164191,165418,152003,202210,181411,
                                    161799,197371,166914,172559,215343,179135,
                                    265269,180518,147237,214652,227319,103188,
                                    183866,222202,191383,191890,109873,
                                    133792,109860,156517,74715,69479,120718,
                                    68687,45030,37129,60886,62786,31732,28295,
                                    32148,40005,14809,11468,17749,17135,
                                    13005,6799,7717,9718,4810,3285,4249,
                                    3036,3287,1759,2402,1738,1010,2177,2063,
                                    1623,1730,1248,1895,2463,2276,3589,
                                    4195,2823,3450,4157,4570,2719,4083,6586,
                                    4617,5137,7796,6564,7405,7298,7867,
                                    7580,9771,11647,25827,25616,15632,10454,
                                    13278,16858,27550,18719,48277,28639,
                                    32971,20762,17972,18975,15609,18617,6124,
                                    2116,3044)
)
```

```
head(cdc)
```

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

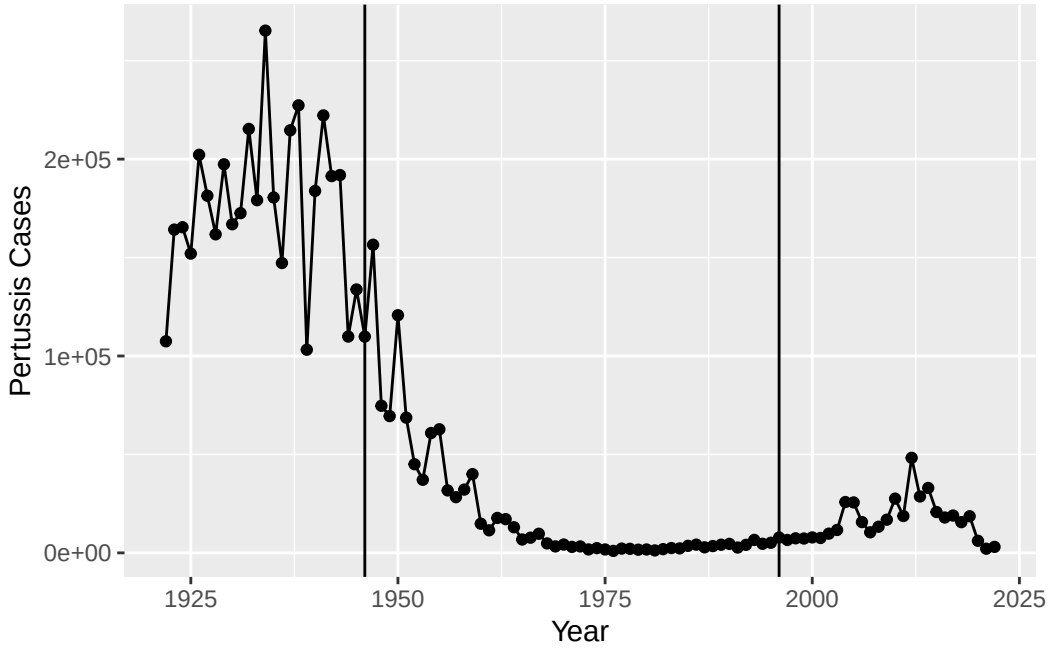
```
library(ggplot2)

ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x = 'Year', y = 'Pertussis Cases')
```



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?

```
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x = 'Year', y = 'Pertussis Cases') +
  geom_vline(xintercept=1996) + geom_vline(xintercept=1946)
```



In 1946, the introduction of the wP vaccine drastically decreased the number of reported Pertussis Cases but in 1996, when the aP vaccine was impletment, cases seemingly increased (not by much however).

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

Cases have seemed to started increasing again. This could be due to increase quality and availibility of testing, a weaker vaccine, or even increased vaccine resistance of viruses/bacteria.

3. Exploring CMI-PB data

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

```
Warning: package 'jsonlite' was built under R version 4.4.3
```

Let's now read the main subject database table from the CMI-PB API. You can find out more about the content and format of this and other tables here: https://www.cmi-pb.org/blog/understand-data/

```
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
87 85
```

or

```
sum(subject[,2] == 'wP')
```

```
[1] 85
```

```
sum(subject[,2] == 'aP')
```

```
[1] 87
```

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

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

```
Female   Male
   112     60
```

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

```r
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
df_summary <- subject %>% count(race, biological_sex)

print(df_summary)
```

```
                                    race biological_sex  n
1               American Indian/Alaska Native           Male  1
2                                      Asian         Female 32
3                                      Asian           Male 12
4                   Black or African American         Female  2
5                   Black or African American           Male  3
6                          More Than One Race         Female 15
7                          More Than One Race           Male  4
8  Native Hawaiian or Other Pacific Islander         Female  1
9  Native Hawaiian or Other Pacific Islander           Male  1
10                   Unknown or Not Reported         Female 14
11                   Unknown or Not Reported           Male  7
12                                     White         Female 48
13                                     White           Male 32
```

or

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

```
        American Indian/Alaska Native Asian Black or African American
  Female                             0    32                         2
```

```
  Male                                1    12                           3

       More Than One Race Native Hawaiian or Other Pacific Islander
  Female                 15                                          1
  Male                    4                                          1

       Unknown or Not Reported White
  Female                      14    48
  Male                         7    32
```

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

Side-Note: Working with dates

```
today()
```

```
[1] "2025-03-12"
```

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

```
Time difference of 9202 days
```

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

```
[1] 25.1937
```

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)

head(subject$age)
```

```
Time differences in days
[1] 14315 20890 15411 13585 12489 13585
```

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

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     22      26      27      27      28      34
```

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     22      32      34      36      39      57
```

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

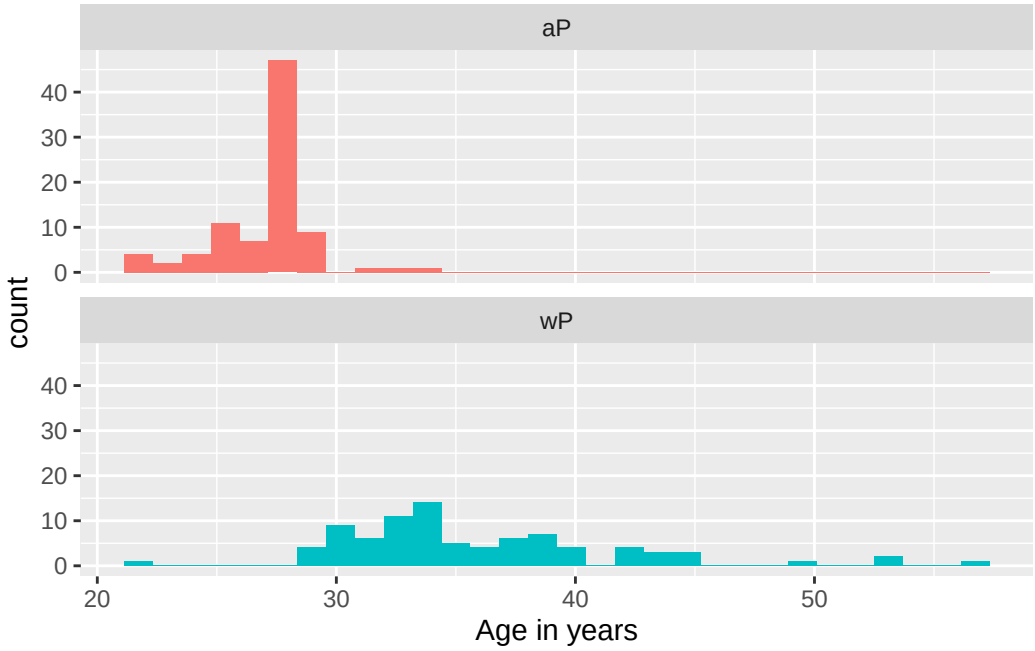
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

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

Q9. With the help of a faceted boxplot 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`.
```



Yes, The graphs clearly illustrate a difference in ages. wP seems to have higher counts in later years while aP has its higher counts much earlier on.

Joining multiple tables

```
# Complete the API URLs...
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:

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

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

```
dim(meta)
```

```
[1] 1503   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 14315 days
2 14315 days
3 14315 days
4 14315 days
5 14315 days
6 14315 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)
```

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

```
dim(abdata)
```

```
[1] 52576    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  5389 10117 10124 10124 10124
```

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$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset 2023_dataset
       31520         8085         7301         5670
```

The data seems to vastly decrease since 2020

4. Examine IgG Ab titer levels

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

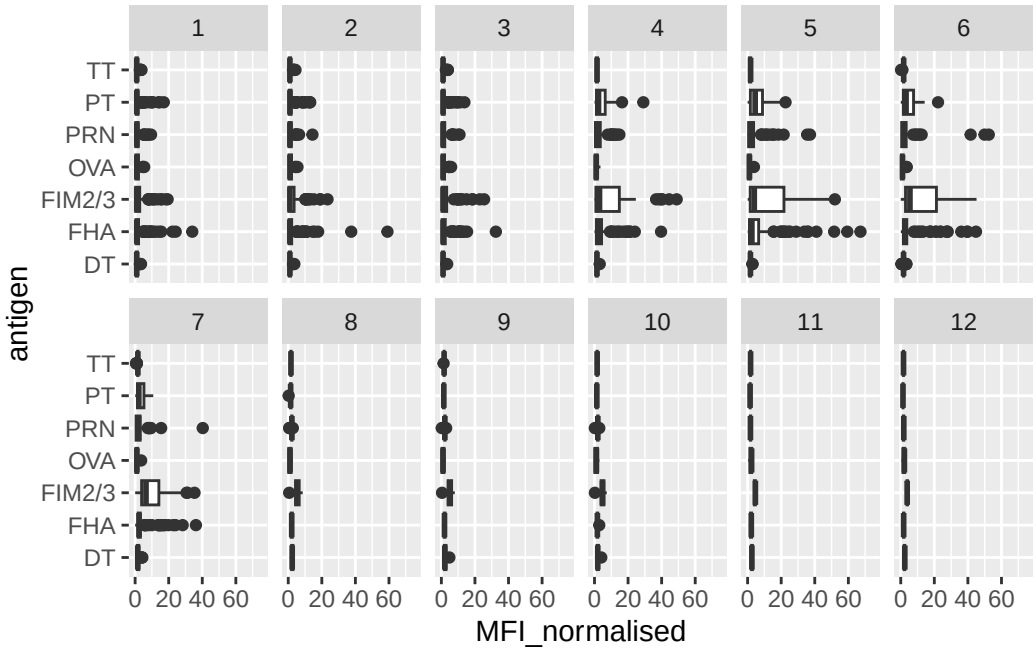
```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
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 14315 days
2 14315 days
3 14315 days
4 15411 days
5 15411 days
6 15411 days
```

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() +
    xlim(0,75) +
  facet_wrap(vars(visit), nrow=2)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

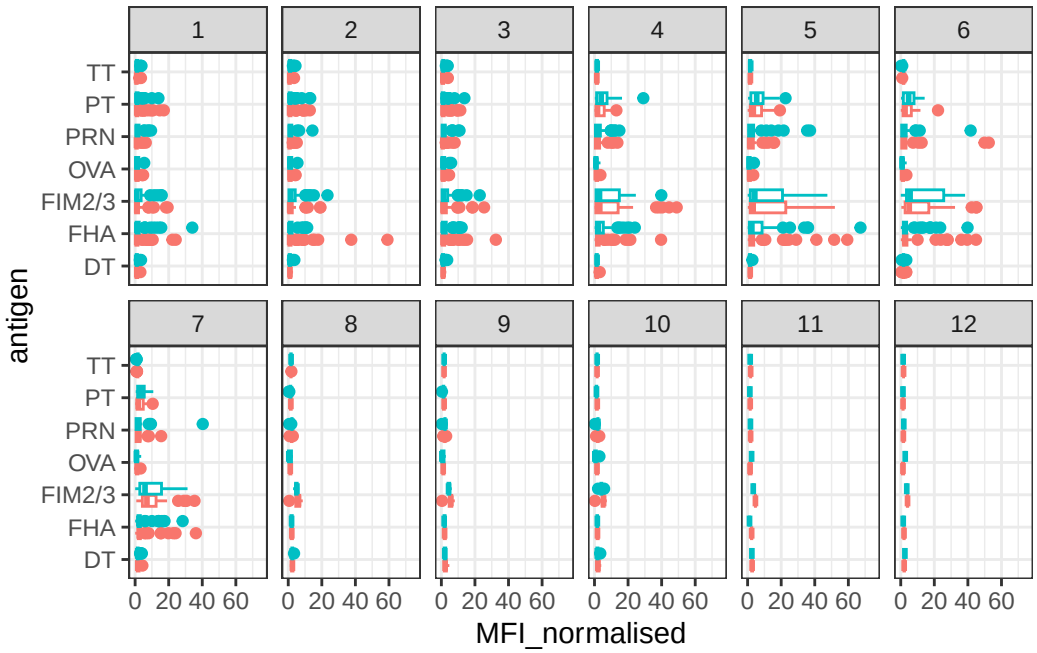


Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

FIM2/3 seem ot have th ehighest levels of titer values across every plot. FIM2/3 is a specific antigen in Bordetall pertussis so it could have a stronger affinity.

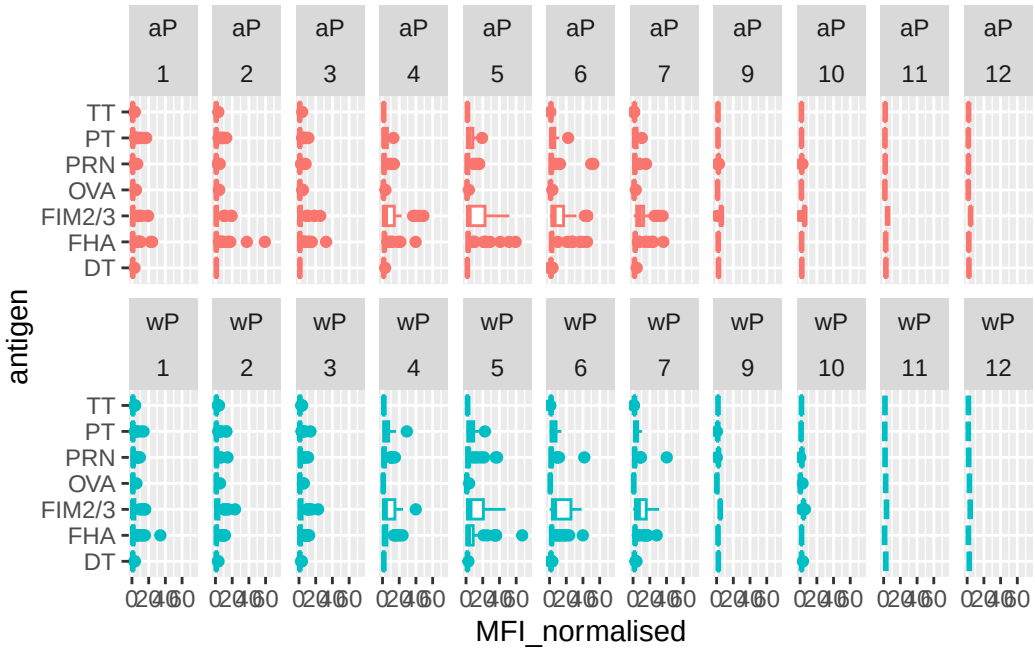
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



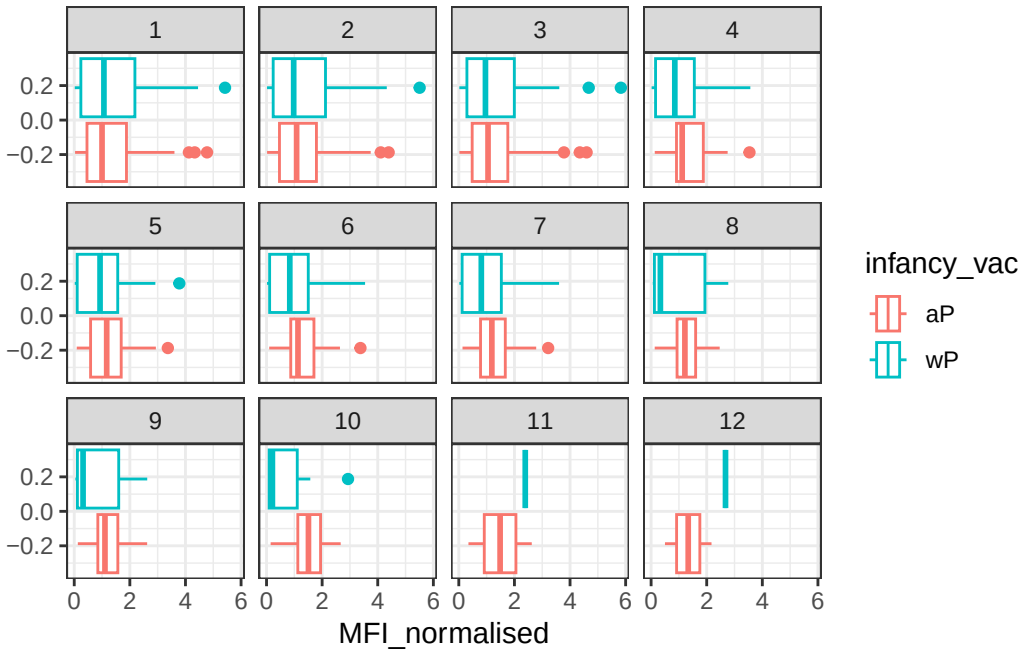

```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



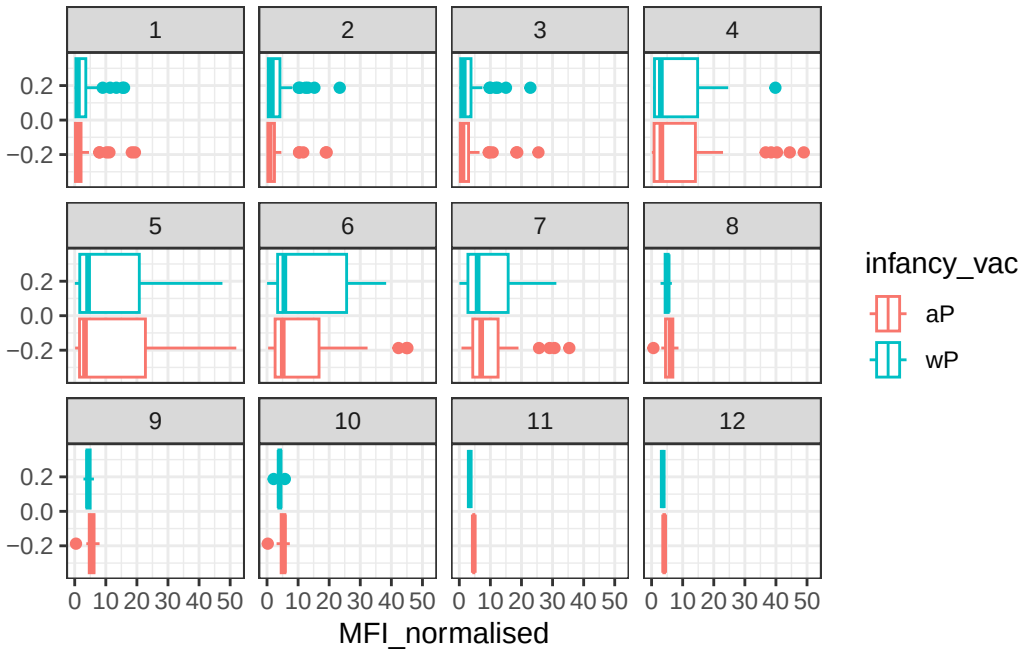
Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium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

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



and the same for antigen=="FIM2/3"

```
filter(igg, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



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

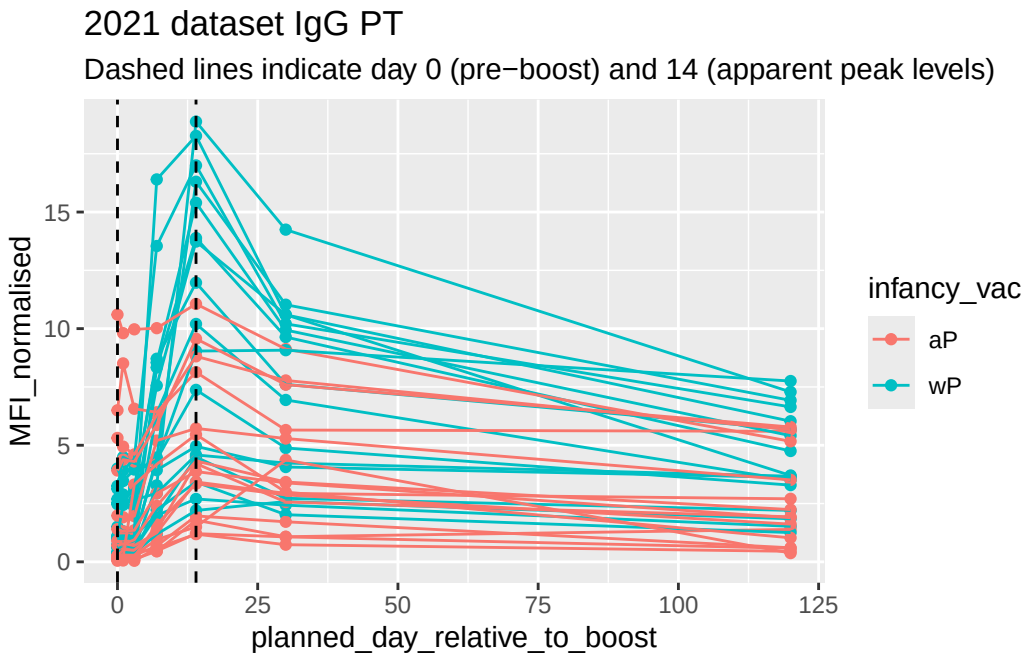
PT levels are rising porportional with time and surpases OVA. They peak at visit 5 and then start to decline for both wP and aP

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

No, there is no clear differences in responses as aP and wP have similar medians across the board with overlapping margins.

```
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?

Here, wP far exceeds aP as MFI levels are much higher in wP.

5. Obtaining CMI-PB RNASeq data

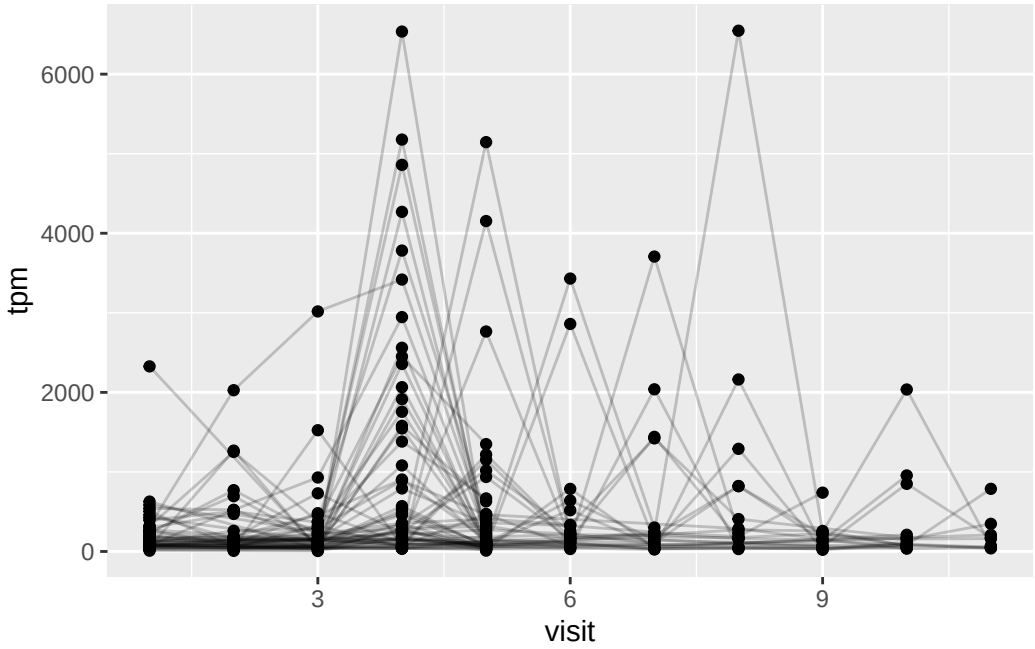
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

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)
```



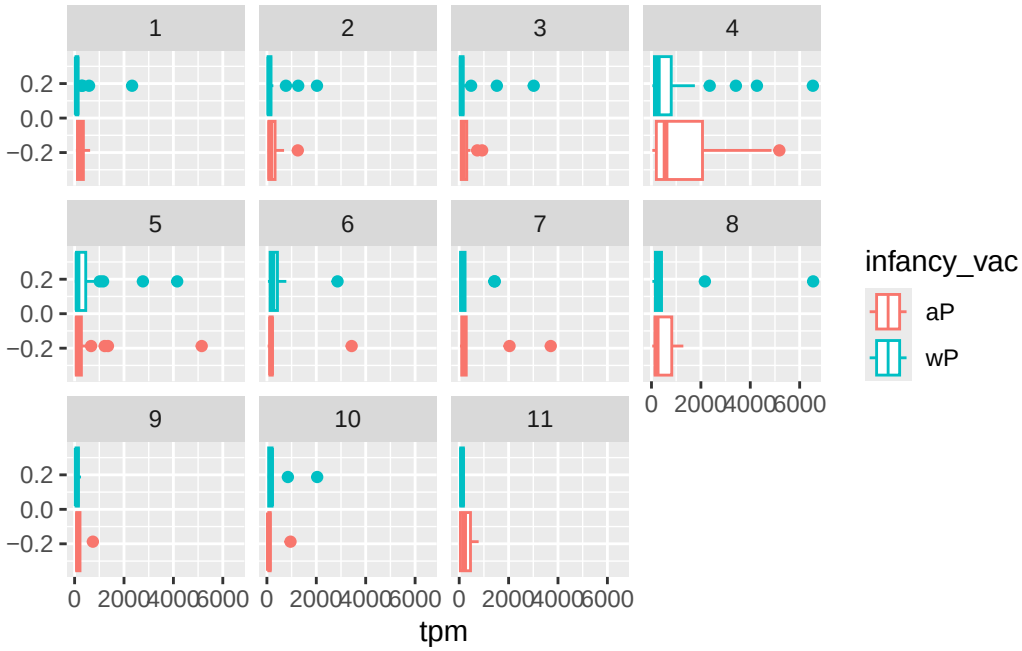
Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

Expression peaks at visit 4 and starts to decrease after.

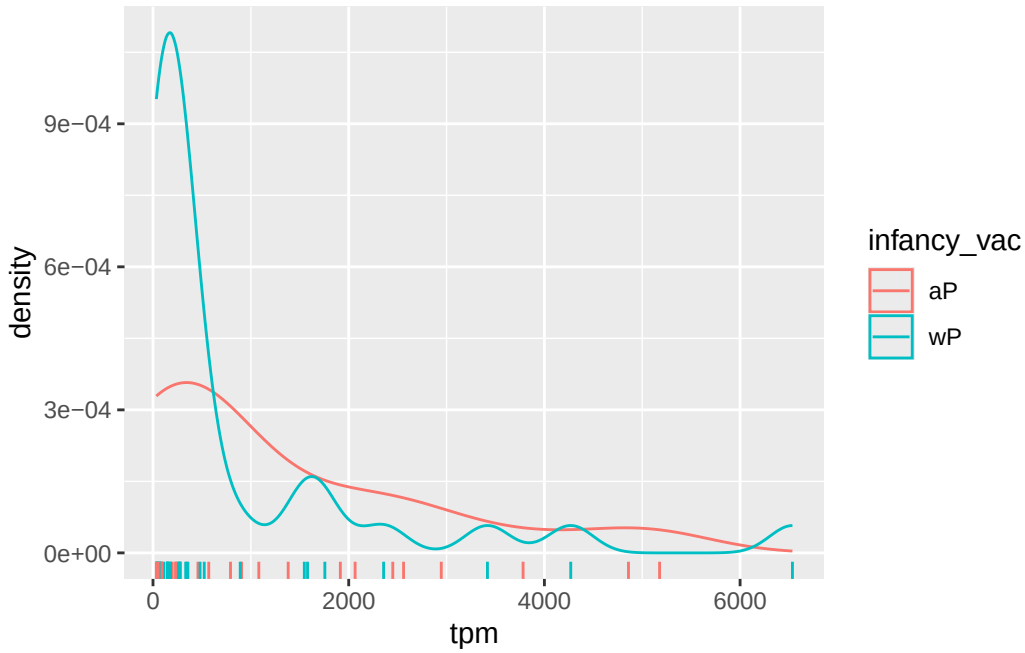
Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

It is similar to the antibody titer data as expressions peak around visit 4/5 but the titer data shows longer expression past visit 7.

```
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()
```


density
9e−04
6e−04
3e−04
0e+00
0
2000
4000
6000
tpm
infancy_vac
aP
wP