Class19: Pertussis and the CMI-PB project

Wanning Cui

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.2
```

```
Warning: package 'readr' was built under R version 4.3.2
```

```
Warning: package 'dplyr' was built under R version 4.3.2
```

```
Warning: package 'stringr' was built under R version 4.3.2
```

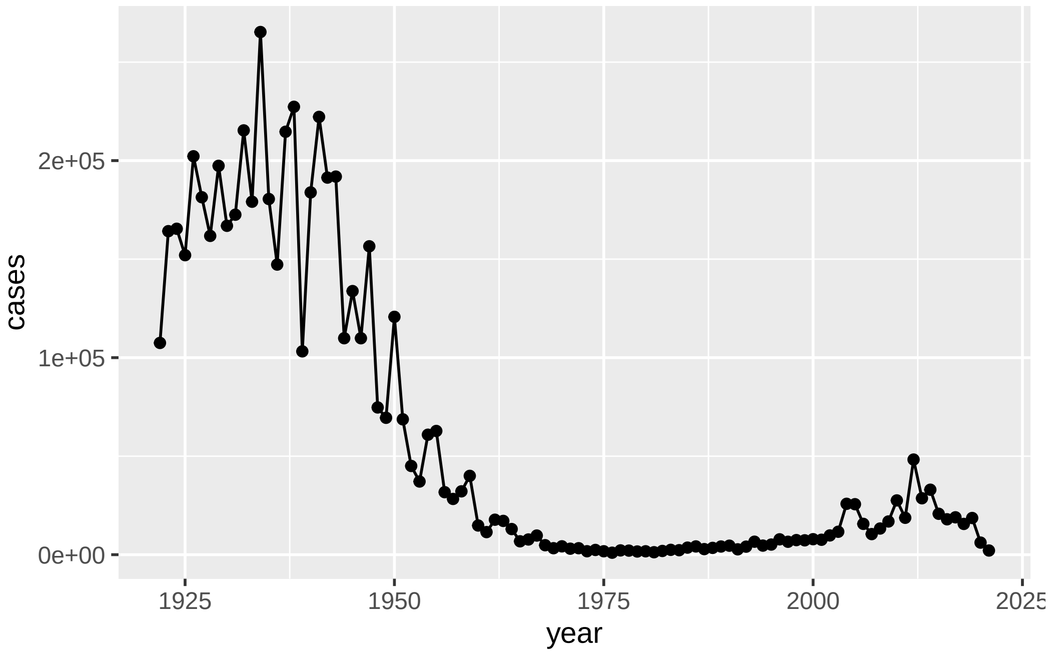
```
Warning: package 'forcats' was built under R version 4.3.2
```

```
Warning: package 'lubridate' was built under R version 4.3.2
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becon
```

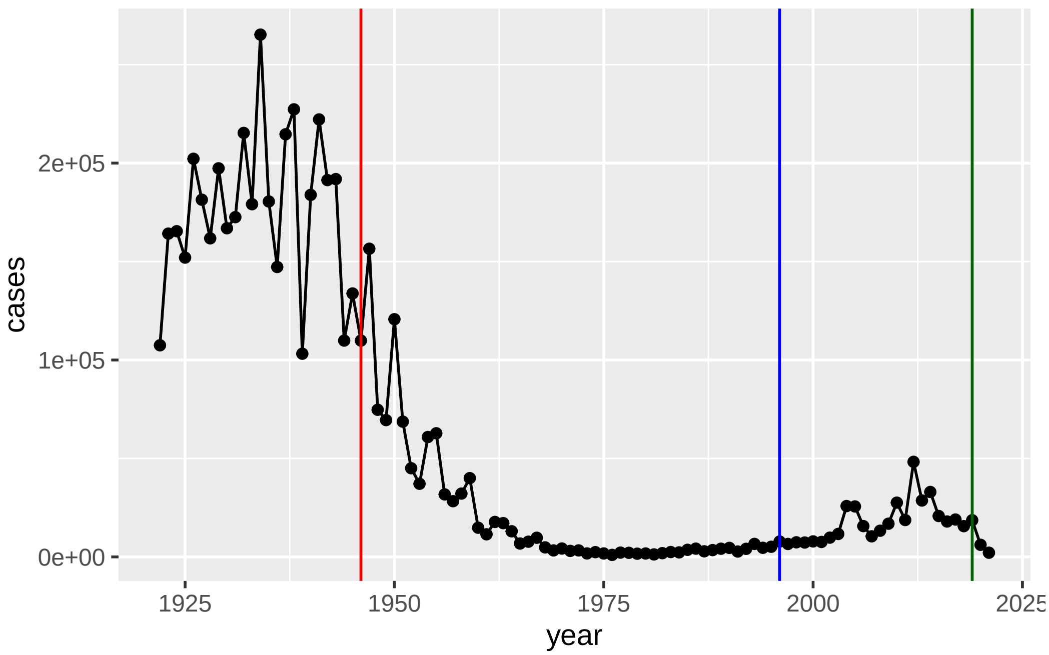
```
#|echo:FALSE
cdc <- data.frame(year = c(1922L,1923L,1924L,1925L,
                                          1926L,1927L,1928L,1929L,1930L,1931L,
                                          1932L,1933L,1934L,1935L,1936L,
                                          1937L,1938L,1939L,1940L,1941L,1942L,
                                          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),
         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)
       )
```

```
ggplot(cdc) +
  aes(x=year, y=cases) +
  geom_point() +
  geom_line()
```



Q2.

```
ggplot(cdc) +
  aes(x=year, y=cases) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept=1946,color="red")+
  geom_vline(xintercept=1996,color="blue")+
  geom_vline(xintercept=2019,color="darkgreen")
```



Q3.

The pertussis cases increase again after the introduction of the vaccine (e.g. around year 2012). This could possibly be due to protests against vaccination or development of more sensitive antibodies.

#CMI-PB project

The CMI-PB project collects and makes available data on the immune response to Pertussis booster vaccination.

We will access this data via the API. We will use the **jsonlite** package to access the data using the `read_json()` function.

```
library(jsonlite)
```

```
Warning: package 'jsonlite' was built under R version 4.3.2
```

```

Attaching package: 'jsonlite'
```

```
The following object is masked from 'package:purrr':

    flatten
```

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

```r
head(subject)
```

```
  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
4          4          wP           Male Not Hispanic or Latino Asian
5          5          wP           Male Not Hispanic or Latino Asian
6          6          wP         Female Not Hispanic or Latino 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
4    1988-01-01    2016-08-29 2020_dataset
5    1991-01-01    2016-08-29 2020_dataset
6    1988-01-01    2016-10-10 2020_dataset
```

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

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

```

aP wP 
60 58 
```

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

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

```

Female   Male 
    79     39 
```

> Q6. What is the breakdown of race and biological sex?

```r
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
```

Q. Make a histogram of the subject age distribution and facet by infancy_vac

```
library(lubridate)
```

```
today() - mdy("06-11-2000")
```

```
Time difference of 8584 days
```

```
time_length( today() - mdy("06-11-2000"),  "years")
```

```
[1] 23.50171
```

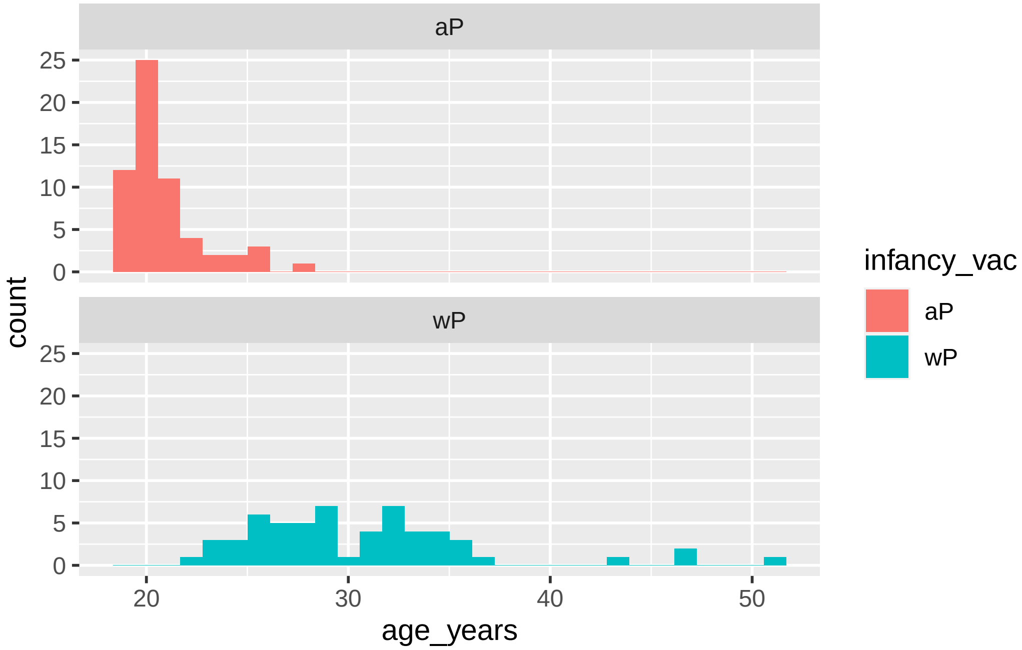
Q7.

```
subject$age <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
```

```
subject$age_years <- time_length(subject$age,"years")
```

```
ggplot(subject)+
  aes(age_years,
      fill=infancy_vac)+
  facet_wrap(vars(infancy_vac),ncol=1)+
  geom_histogram()
```

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



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?

Yes.

There are 3 main datasets in the CMI-PB project at the time of writing.

```
table(subject$dataset)
```

```

2020_dataset 2021_dataset 2022_dataset 
          60           36           22 
```

```r
specimen <- read_json("https://www.cmi-pb.org/api/specimen",simplifyVector=TRUE)
titer <- read_json("https://www.cmi-pb.org/api/v4/plasma_ab_titer",simplifyVector=TRUE)
```

```r
head(specimen)
```

```
  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
1                             0         Blood     1
2                             1         Blood     2
3                             3         Blood     3
4                             7         Blood     4
5                            14         Blood     5
6                            30         Blood     6
```

```r
head(titer)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgE               FALSE   Total 1110.21154       2.493425
2           1     IgE               FALSE   Total 2708.91616       2.493425
3           1     IgG                TRUE      PT   68.56614       3.736992
4           1     IgG                TRUE     PRN  332.12718       2.602350
5           1     IgG                TRUE     FHA 1887.12263      34.050956
6           1     IgE                TRUE     ACT    0.10000       1.000000
   unit lower_limit_of_detection
1 UG/ML                 2.096133
2 IU/ML                29.170000
3 IU/ML                 0.530000
4 IU/ML                 6.205949
5 IU/ML                 4.679535
6 IU/ML                 2.816431
```

Q9. I want to (join) merge the specimen and subject tables together.

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

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

```
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 age_years
1 11212 days  30.69678
2 11212 days  30.69678
3 11212 days  30.69678
4 11212 days  30.69678
5 11212 days  30.69678
6 11212 days  30.69678
```

Q10.

```
abdata <- inner_join(titer, meta)
```

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

```
dim(abdata)
```

```
[1] 41810    22
```

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$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset 
       31520         8085         2205 
```

The number of rows are decreasing.

Q13.

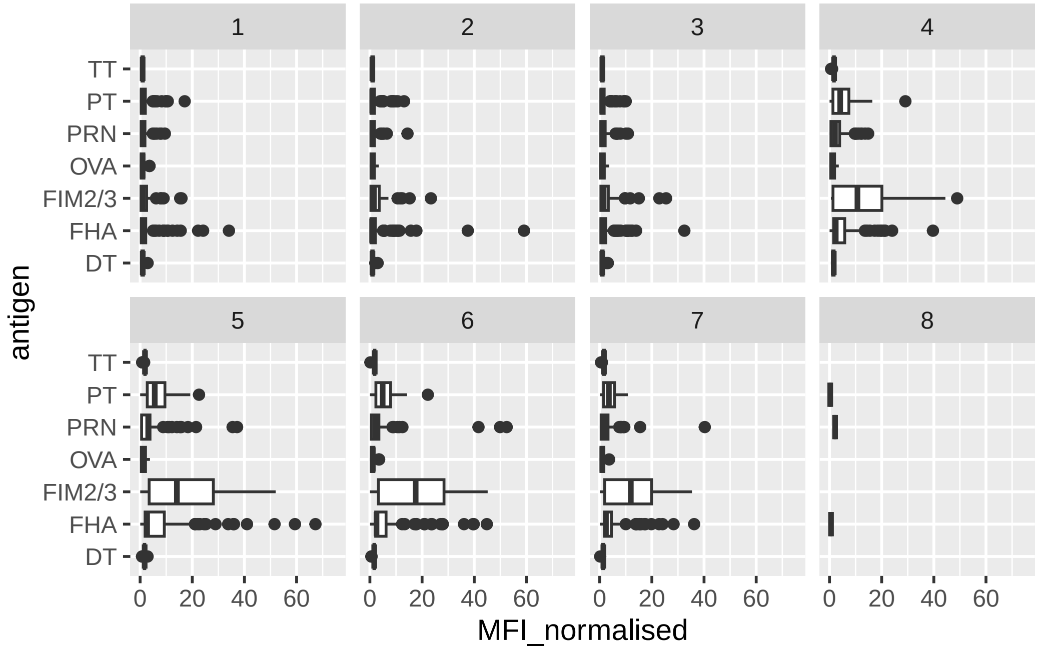
```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

```
  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 age_years
1 11212 days  30.69678
2 11212 days  30.69678
3 11212 days  30.69678
4 12336 days  33.77413
5 12336 days  33.77413
6 12336 days  33.77413
```

```r
ggplot(igg)+
  aes(MFI_normalised,antigen)+
  geom_boxplot()+
  xlim(0,75)+
  facet_wrap(vars(visit),nrow=2)
```

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



```
oops <- abdata %>% filter(antigen == "Fim2/3")
table(oops$dataset)
```

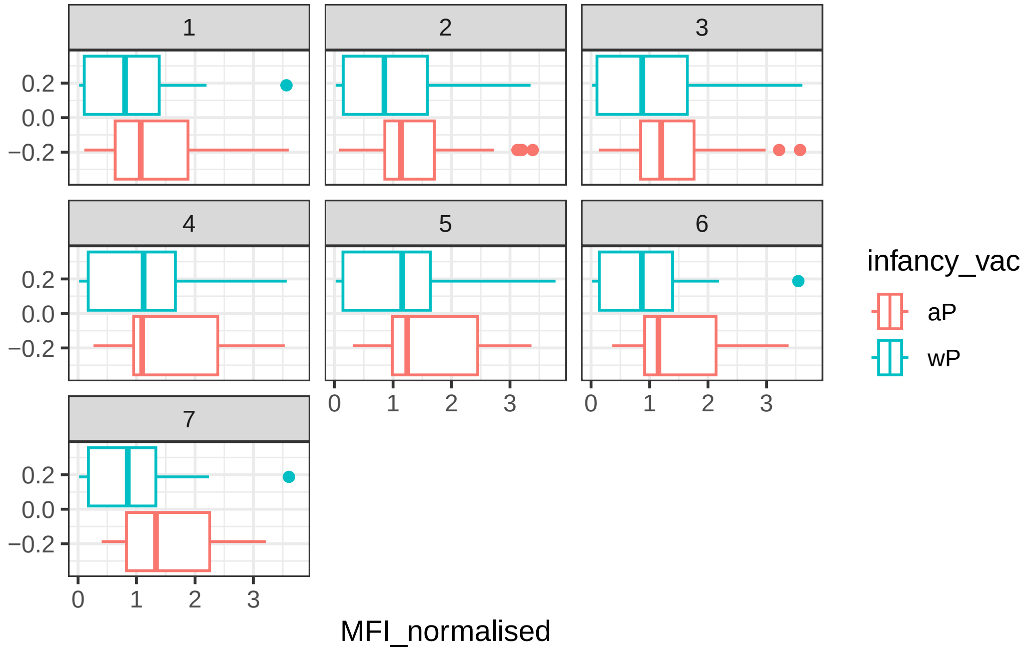
```
< table of extent 0 >
```

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

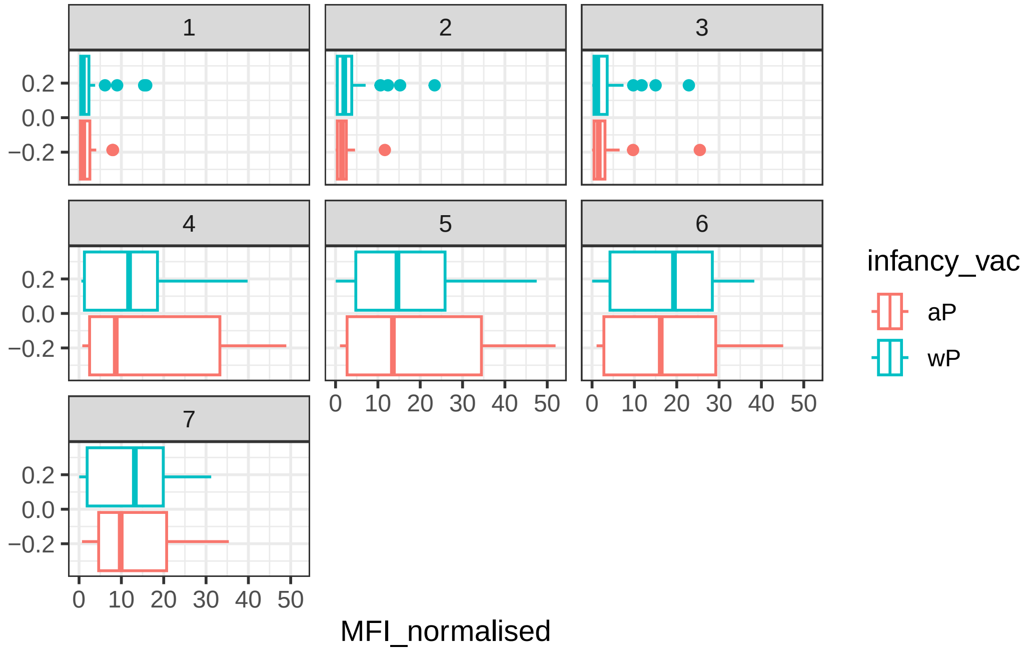
The antigens are Fim2/3 and PT. This is because they are immunogenic.

Q15.

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



```
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 clearly rise over time and far exceed those of OVA. They also appear to peak at visit 5 and then decline. This trend appears similar for wP and aP subjects.

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

The aP responses for OVA antigen is a bit higher than wP responses. However, the wP responses for PT antigen is higher than aP responses.

Select (or filter) for the 2021 dataset and isotype IgG I want a time course (`planned_day_relative_to_boost`) of IgG versus (`MFI_normalised`)for"PT"angigen

```
igpt.21 <- abdata %>% filter (dataset == "2021_dataset",
                              isotype == "IgG",
                              antigen == "PT")
```

```
ggplot(igpt.21)+
  aes(planned_day_relative_to_boost,
      MFI_normalised,
      col=infancy_vac)+
  geom_point()+
```

```r
  geom_line(aes(group=subject_id), linewidth=0.5,alpha=0.5)+
  geom_smooth(se=FALSE,span=0.4,linewidth=3)
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -0.6
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3.6
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 1.8382e-16
```

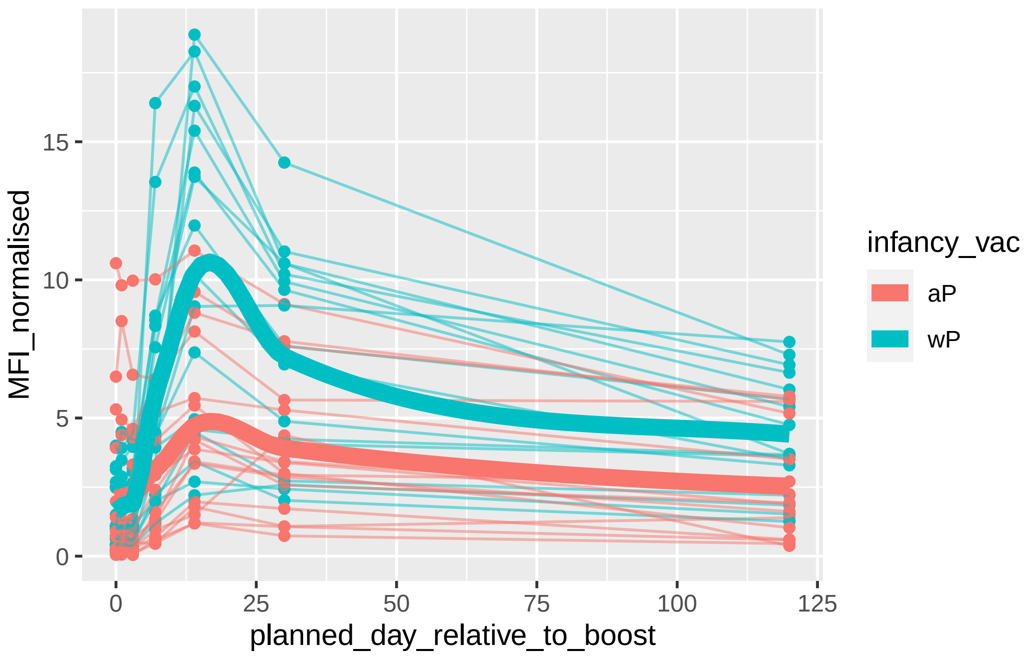
```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 11364
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -0.6
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3.6
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 1.4316e-16
```

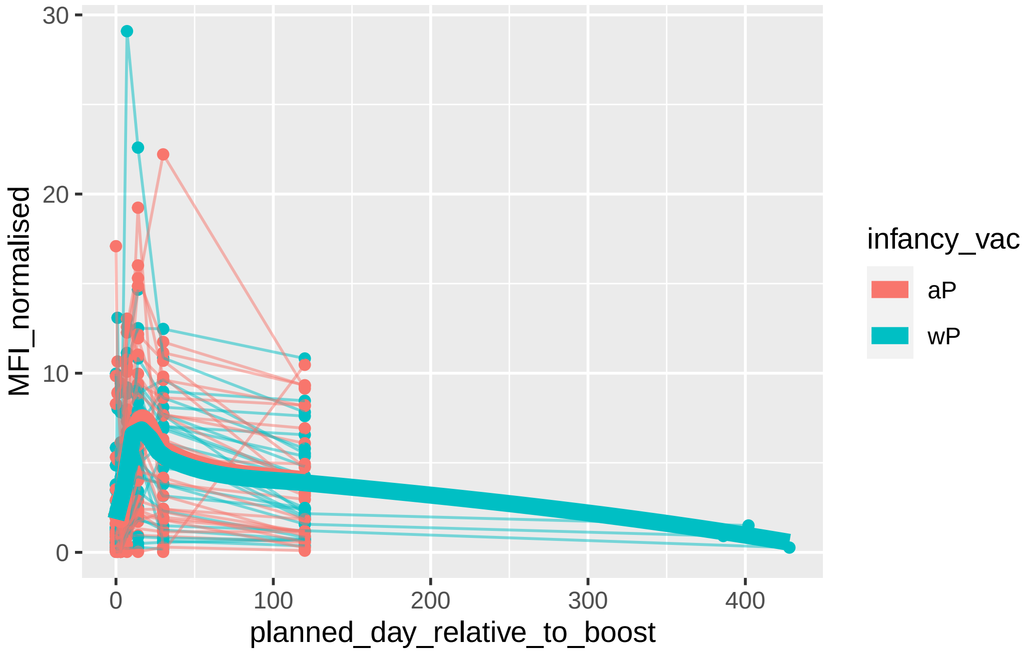
```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 11364
```



```
igpt.21 <- abdata %>% filter (dataset == "2020_dataset",
                              isotype == "IgG",
                              antigen == "PT")
ggplot(igpt.21)+
  aes(planned_day_relative_to_boost,
      MFI_normalised,
      col=infancy_vac)+
  geom_point()+
  geom_line(aes(group=subject_id), linewidth=0.5,alpha=0.5)+
  geom_smooth(se=FALSE,span=0.4,linewidth=3)
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -0.6

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3.6

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 2.9482e-16
```

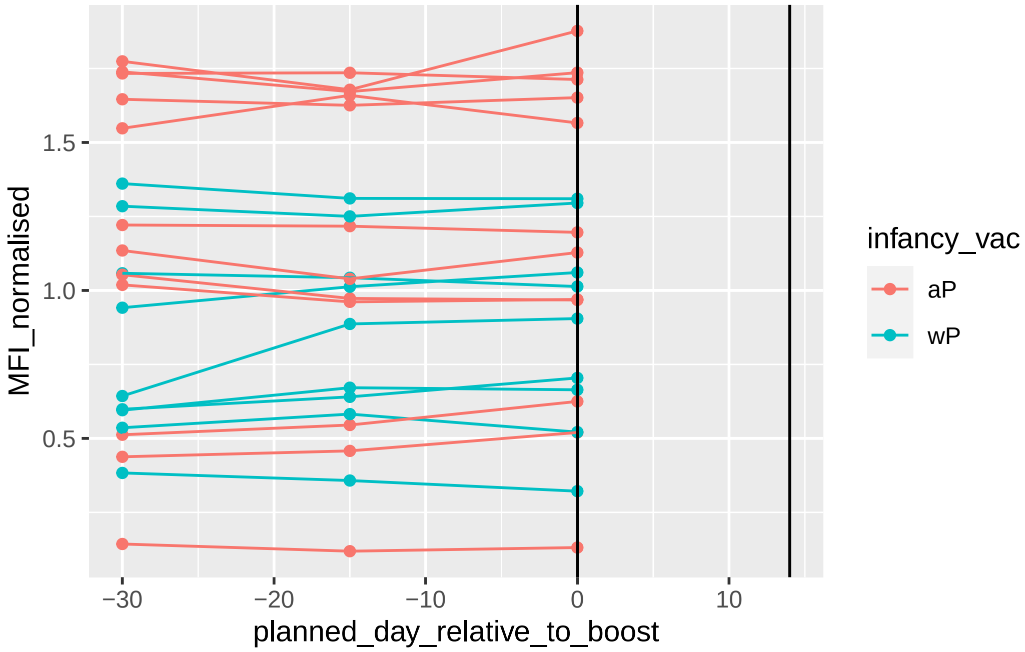
```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 11364

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -2.14

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 5.14

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 4.7594e-16

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 9
```



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

Yes.

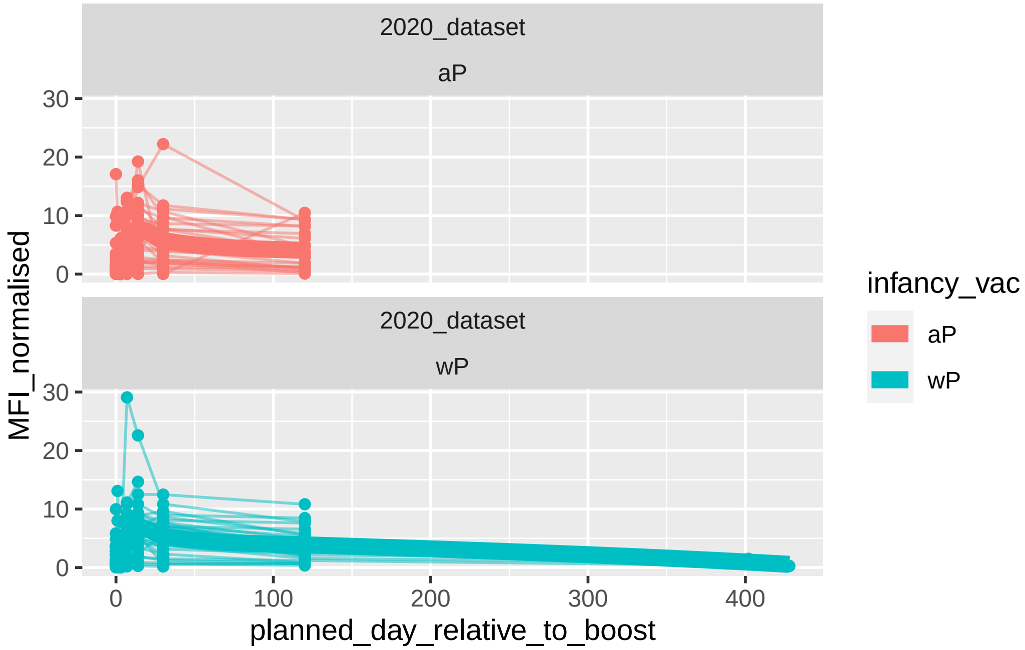
```
igpt.22 <- abdata %>% filter (dataset == "2022_dataset",
                              isotype == "IgG",
                                           antigen == "PT")
ggplot(igpt.22)+
  aes(planned_day_relative_to_boost,
      MFI_normalised,
      col=infancy_vac)+
  geom_point()+
  geom_line(aes(group=subject_id))+
  geom_vline(xintercept=0) +
  geom_vline(xintercept=14)
```



Make a plot of IgG versus MFI_normalised for "PT" antigen and facet by dataset and infancy_vac

```
ggplot(igpt.21)+
  aes(planned_day_relative_to_boost,
      MFI_normalised,
      col=infancy_vac)+
  geom_point()+
  geom_line(aes(group=subject_id), linewidth=0.5,alpha=0.5)+
  geom_smooth(se=FALSE,span=0.4,linewidth=3)+
  facet_wrap(vars(dataset,infancy_vac),ncol=1)
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -0.6

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3.6

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 2.9482e-16

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 11364

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -2.14

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 5.14

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 4.7594e-16

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 9
```



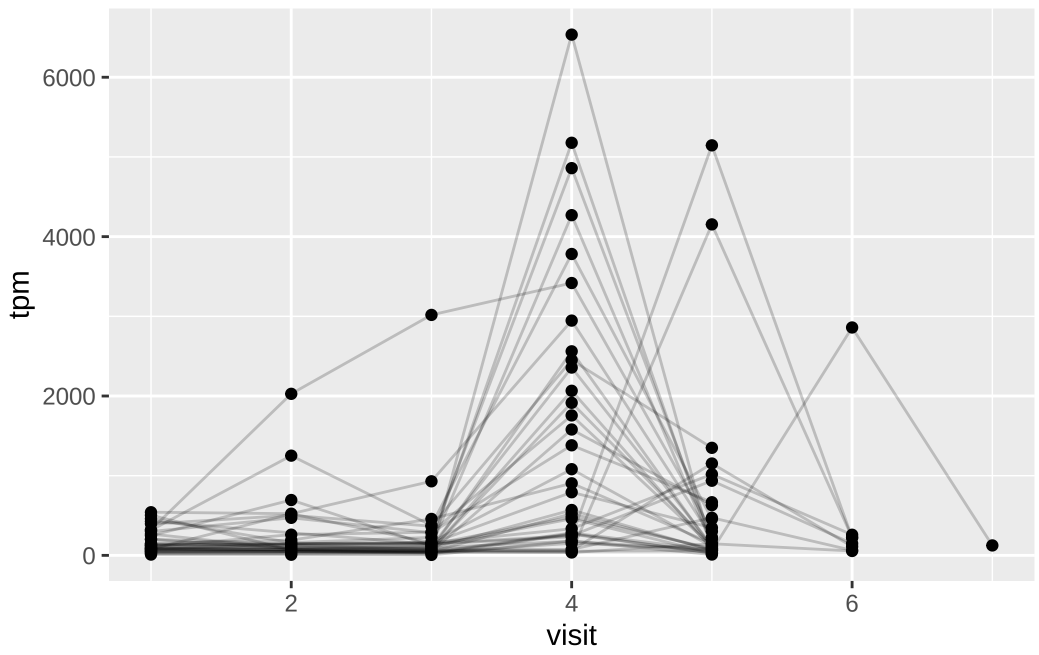
```r
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.
rna <- read_json(url, simplifyVector = TRUE)
```

```r
ssrna <- inner_join(rna, meta)
```

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

Q19.

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



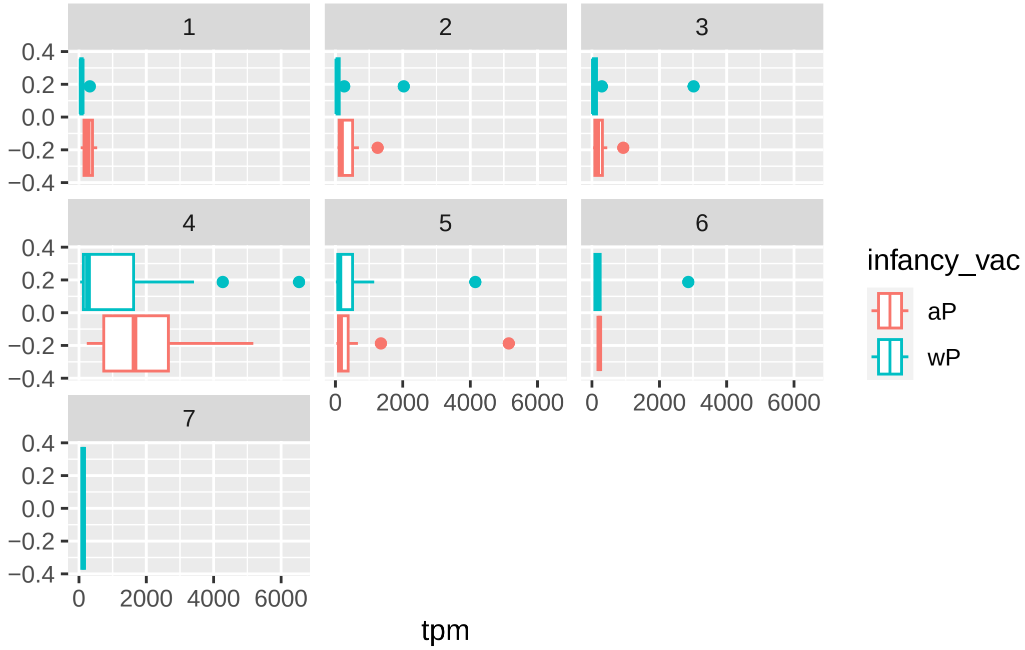
> Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

The gene expression gradually increases and reaches peak at 4th visit.

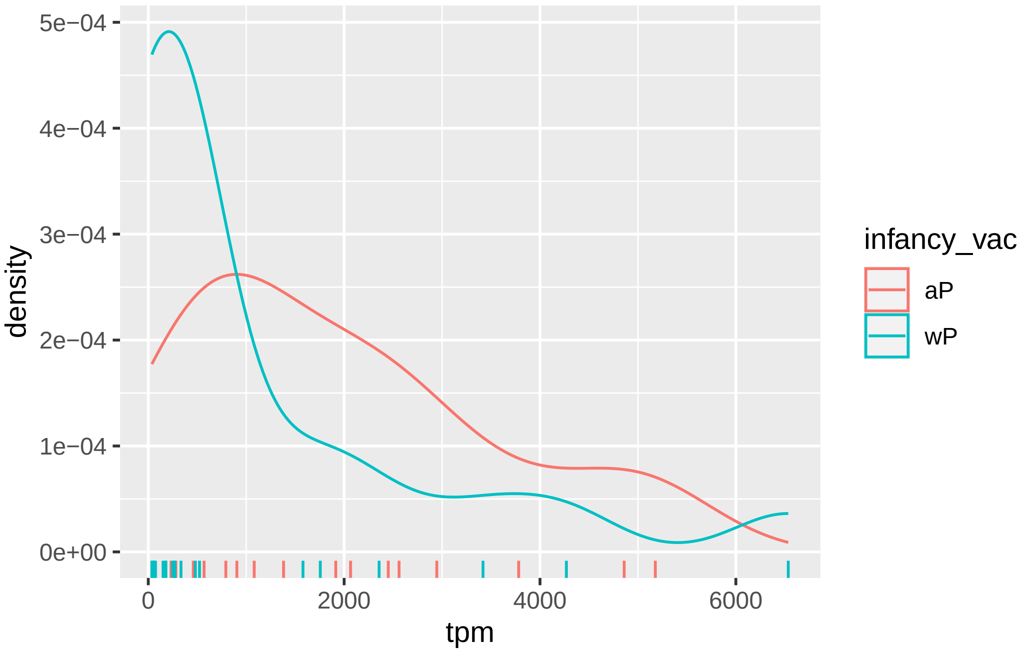
> Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

No. The antibody titer data shows at later timepoints. This is because antibodies made by cells live longer.

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



```r
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