Investigating Pertussis Resurgence

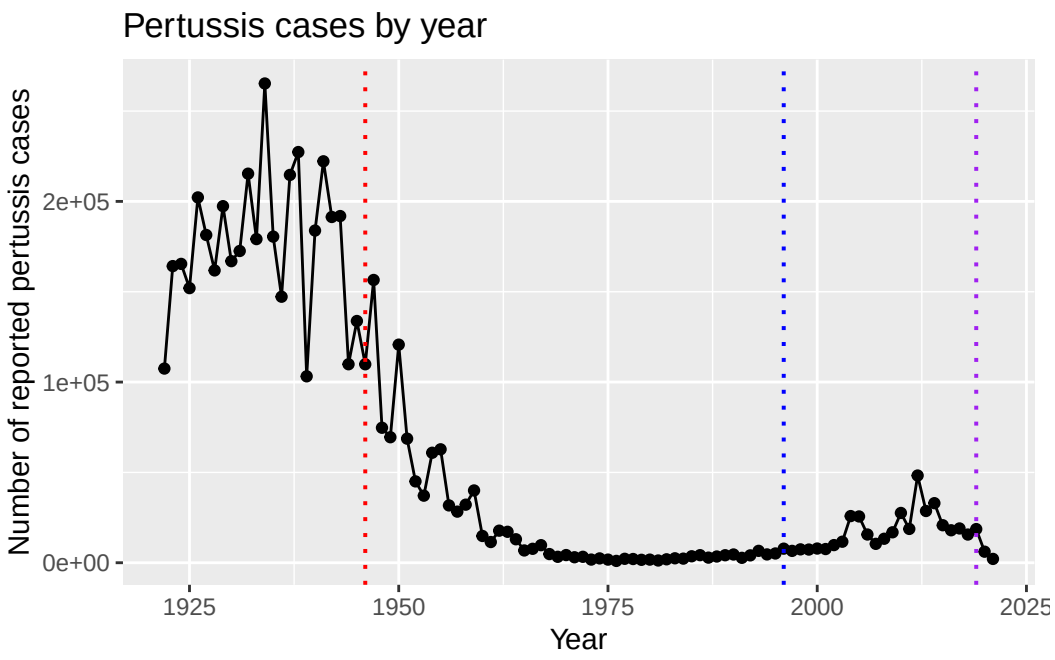
Jazz Zhang A16149005

```
library(ggplot2)

# Q1
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),
     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)
    )

plot <- ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(title = "Pertussis cases by year", y = "Number of reported pertussis cases")
```

```
plot +
  geom_vline(xintercept = c(1946, 1996, 2019), linetype = "dotted", color = c("red", "blue
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Q2. Number of reported pertussis cases reduced dramatically after introduction of wP vaccine

Q3. Number of reported pertussis cases increased after switching to aP vaccine. Protection provided by the aP vaccine wanes faster than the wP vaccine

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

```
aP wP
60 58
```

```
# Q5
table(subject$biological_sex)
```

```
Female   Male
    79     39
```

```
# Q6
table(subject$biological_sex, subject$ethnicity)
```

```
        Hispanic or Latino Not Hispanic or Latino Unknown
  Female                 21                     57       1
  Male                    5                     31       3
```

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2023-12-09"
```

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

round(mean(subject$age[subject$infancy_vac=="wP"]))
```

```
[1] 36
```

```
round(mean(subject$age[subject$infancy_vac=="aP"]))
```

```
[1] 26
```

```
t.test(subject$age[subject$infancy_vac=="wP"], subject$age[subject$infancy_vac=="aP"])$p.v
```

```
[1] 6.813505e-19
```

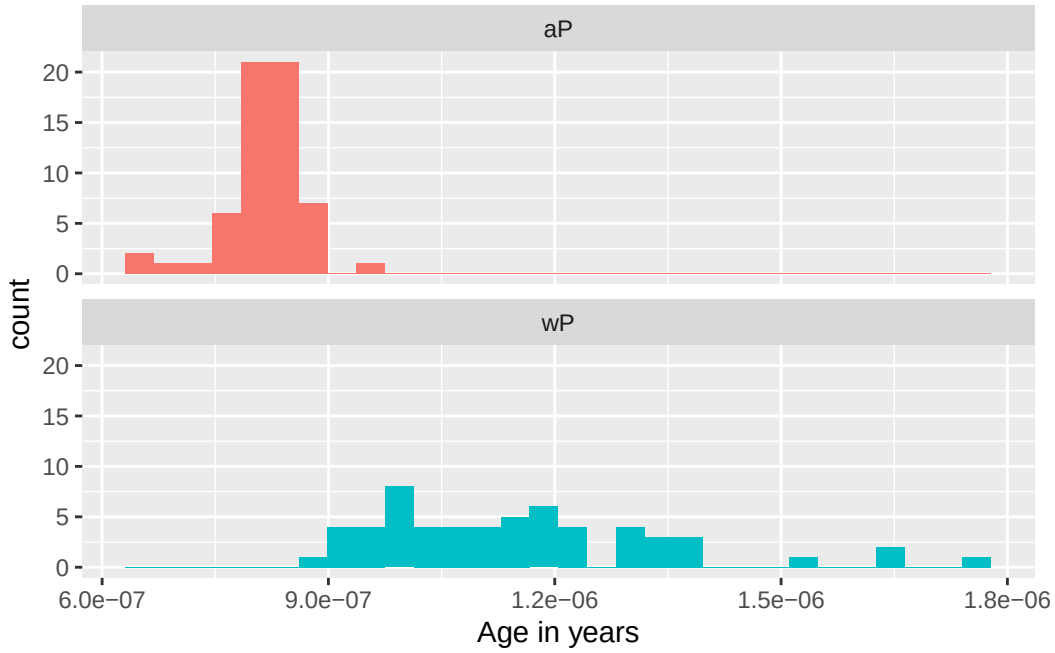
Q7. i) 36; ii) 26; iii) they are significantly different

```
# Q8
subject$age_at_boost <- time_length( ymd(subject$date_of_boost) - ymd(subject$year_of_birt
```

```
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`.
```




```
t.test(subject$age_at_boost[subject$infancy_vac=="wP"], subject$age_at_boost[subject$infan
```

```
[1] 9.121472e-19
```

Q9. They are significantly different

```
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)
```

```
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
```

```
# Q9
meta <- inner_join(specimen, subject)
```

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

```
dim(meta)
```

```
[1] 939  15
```

```
# Q10
abdata <- inner_join(titer, meta)
```

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

```
dim(abdata)
```

```
[1] 41810    22
```

```
# Q11
table(abdata$isotype)
```

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

```r
# Q12
table(abdata$dataset)
```

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

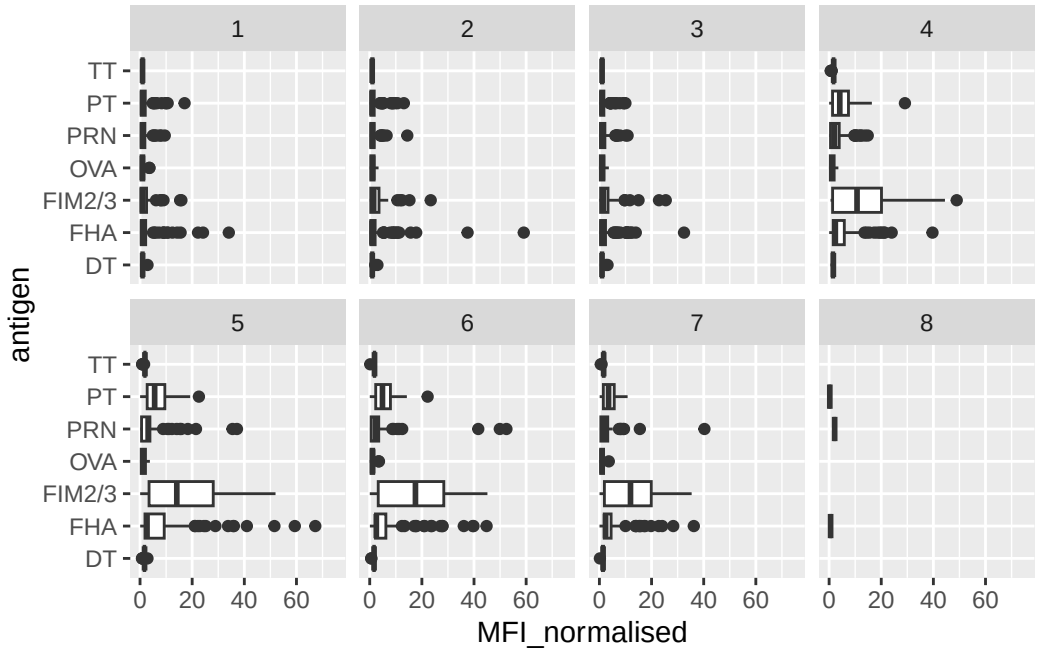
```r
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

```
  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_at_boost
1 37.93566     30.69678
```

```
2 37.93566     30.69678
3 37.93566     30.69678
4 40.93634     33.77413
5 40.93634     33.77413
6 40.93634     33.77413
```

```r
# Q13
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
    xlim(0,75) +
  facet_wrap(vars(visit), nrow=2)
```

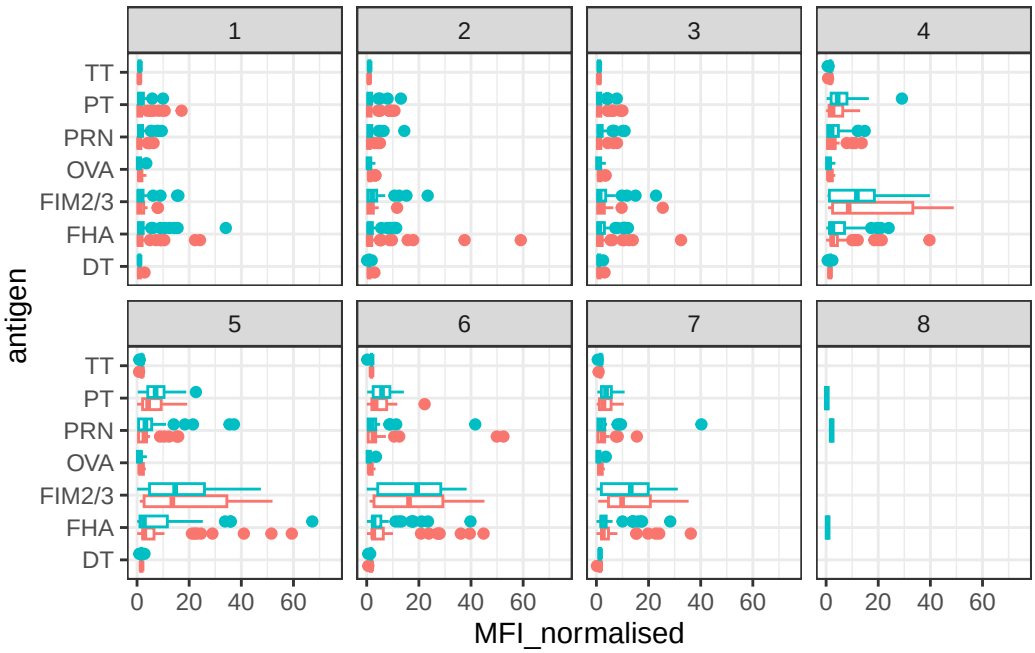
```
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).
```



Q14. Pertussis toxin, pertactin, fimbriae, and filamentous hemagglutinin show differences in the level of IgG antibody over time. These are surface molecules of pertussis which the patients are vaccinated against. Tetanus toxoid and diphtheria toxin are toxins of a different species of bacteria, while OVA albumin is a control peptide. Patients receiving pertussis vaccine should not generate antibodies against these molecules.

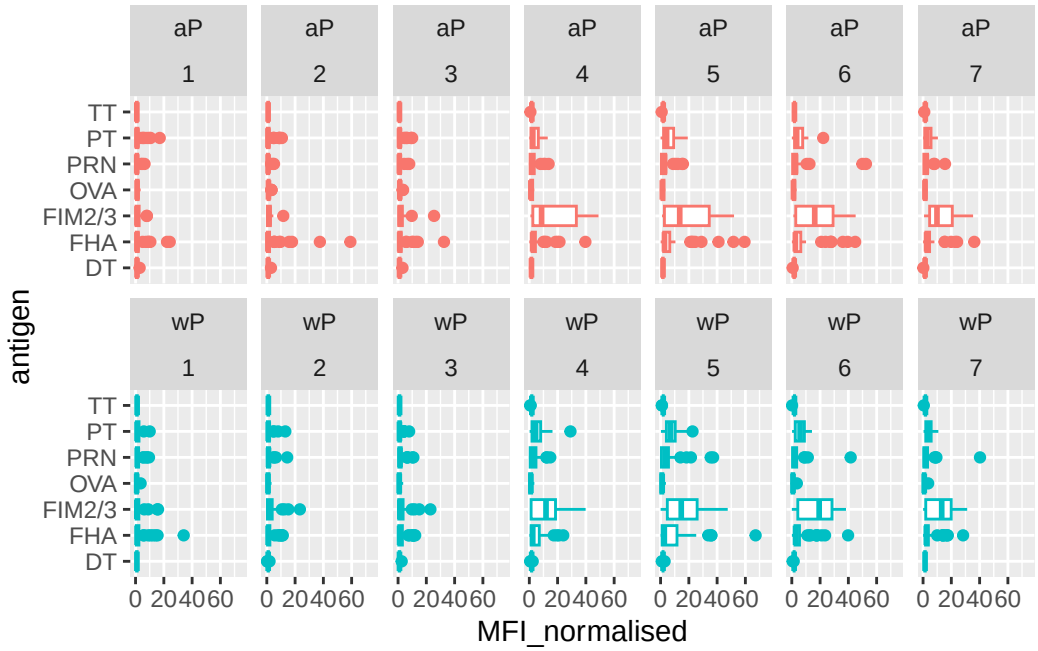
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

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

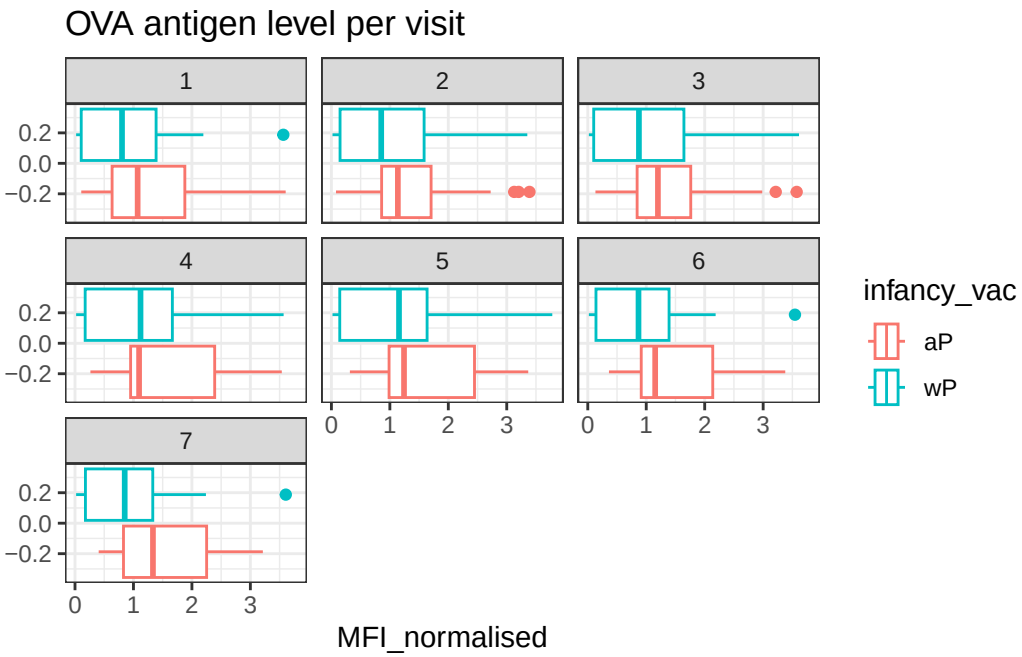



```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

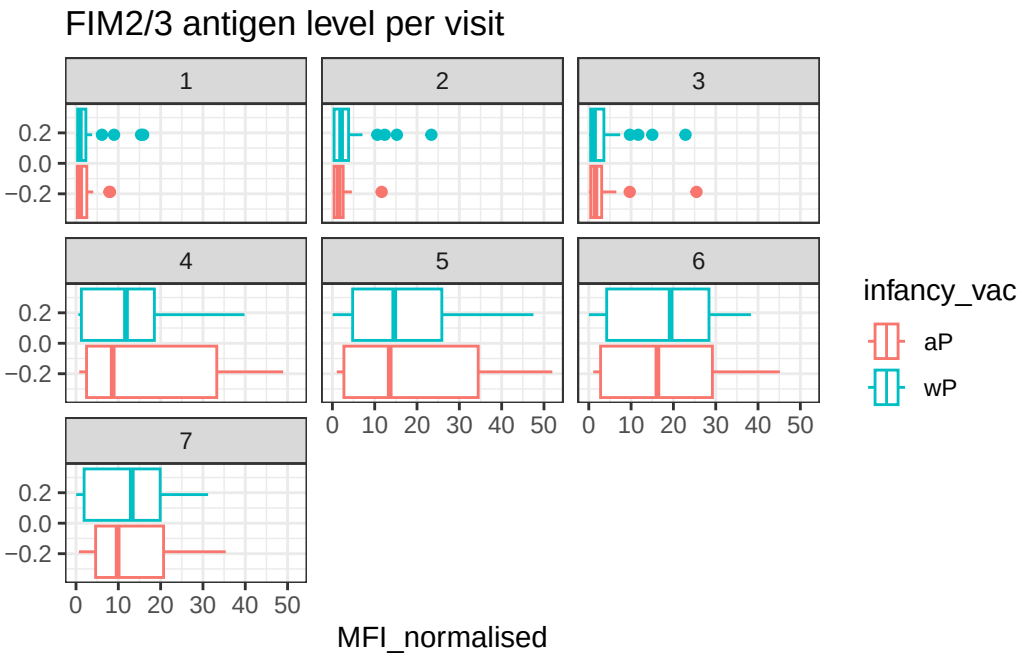
```
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).
```




```
# Q15
filter(igg, antigen=="OVA") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title = "OVA antigen level per visit")
```



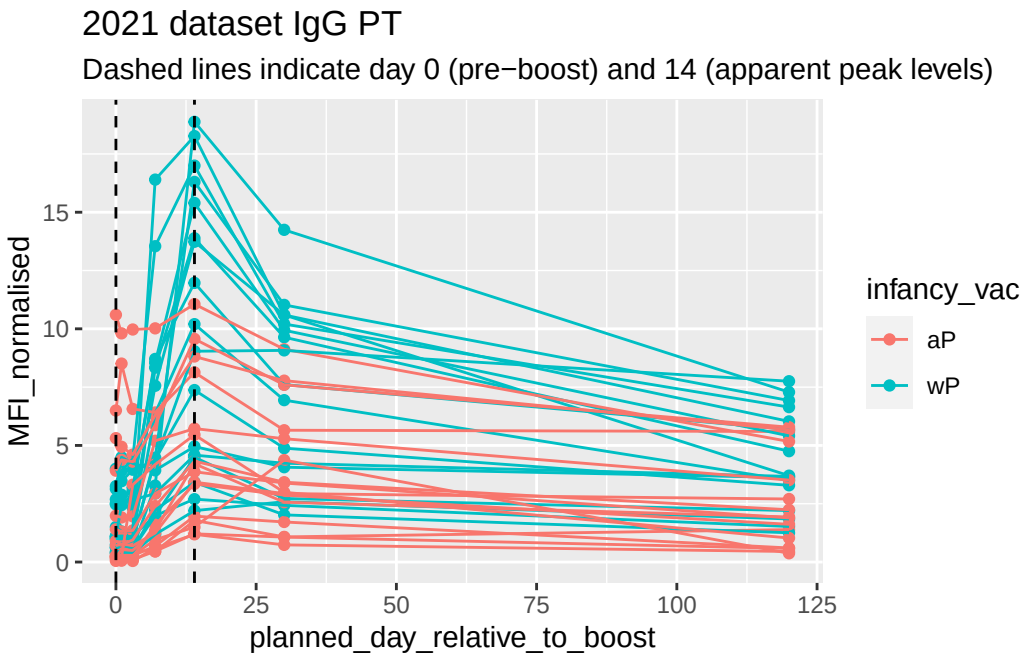

```
filter(igg, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title = "FIM2/3 antigen level per visit")
```



Q16. IgG level against FIM2/3 increased over time but not against OVA. The trend is similar between aP and wP responses.

Q17. There is no clear difference between aP and wP responses

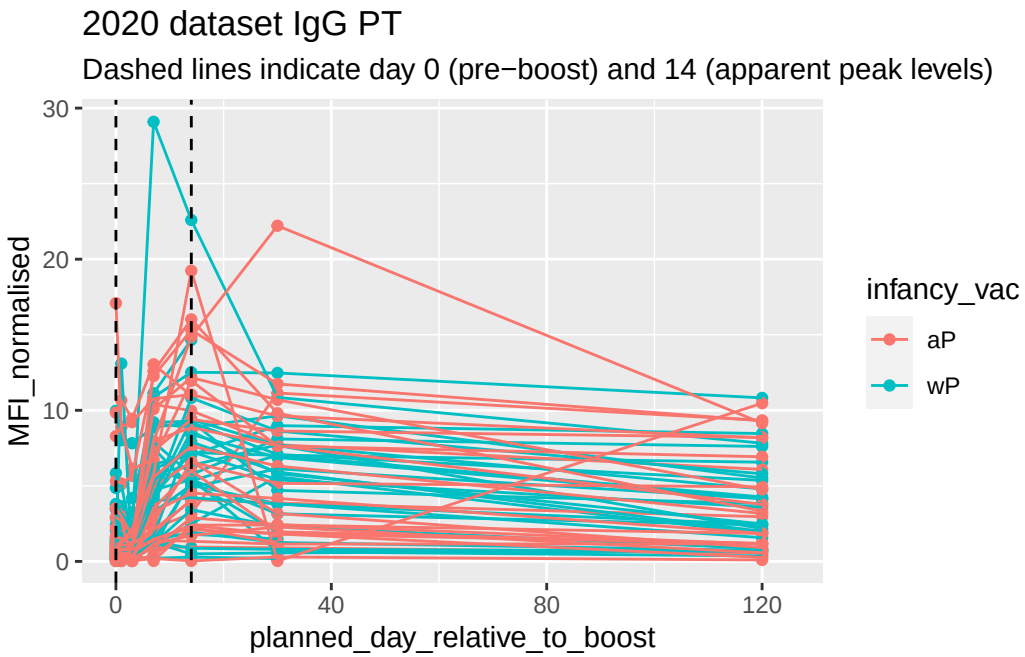
```
abdata.21 <- abdata %>% filter(dataset == "2021_dataset")

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)")
```




```
abdata.20 <- abdata %>% filter(dataset == "2020_dataset")

abdata.20 %>%
  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="2020 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)") +
  xlim(0, 125)
```

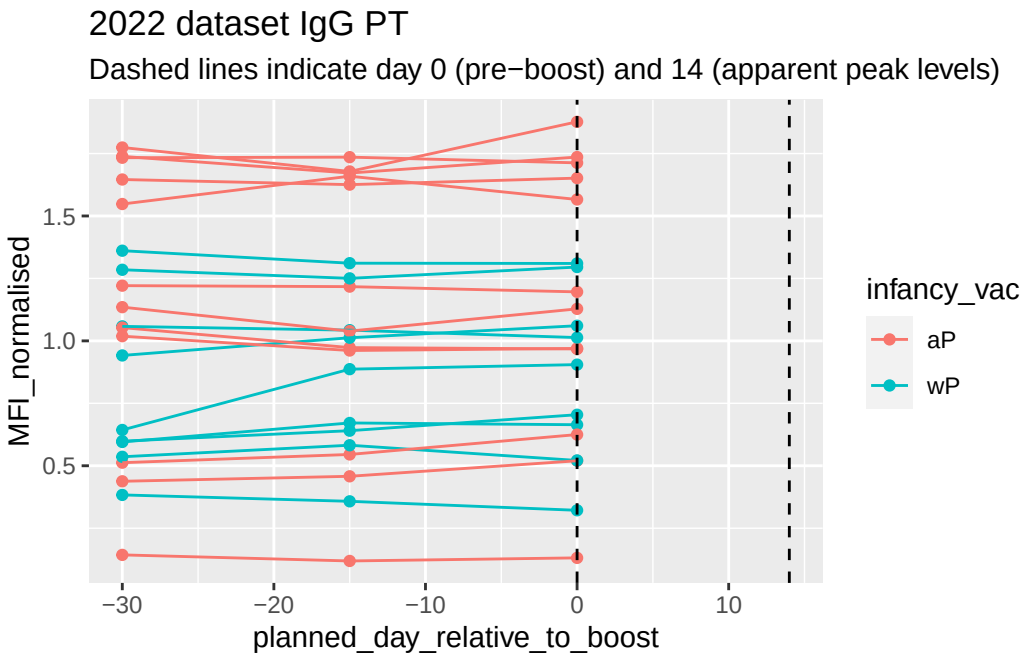
```
Warning: Removed 3 rows containing missing values (`geom_point()`).
```

```
Warning: Removed 3 rows containing missing values (`geom_line()`).
```



Q18. They don't look the same, the 2021 dataset show wP response show higher antibody titer on day 14 but not in 2020 dataset

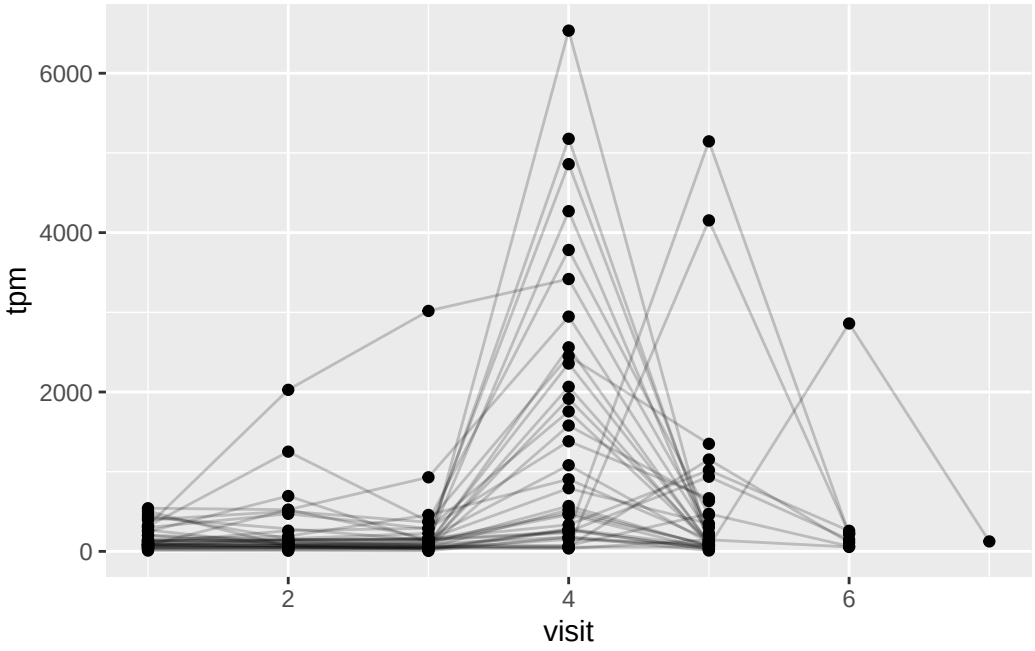
```
abdata.22 <- abdata %>% filter(dataset == "2022_dataset")

abdata.22 %>%
  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="2022 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```




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

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

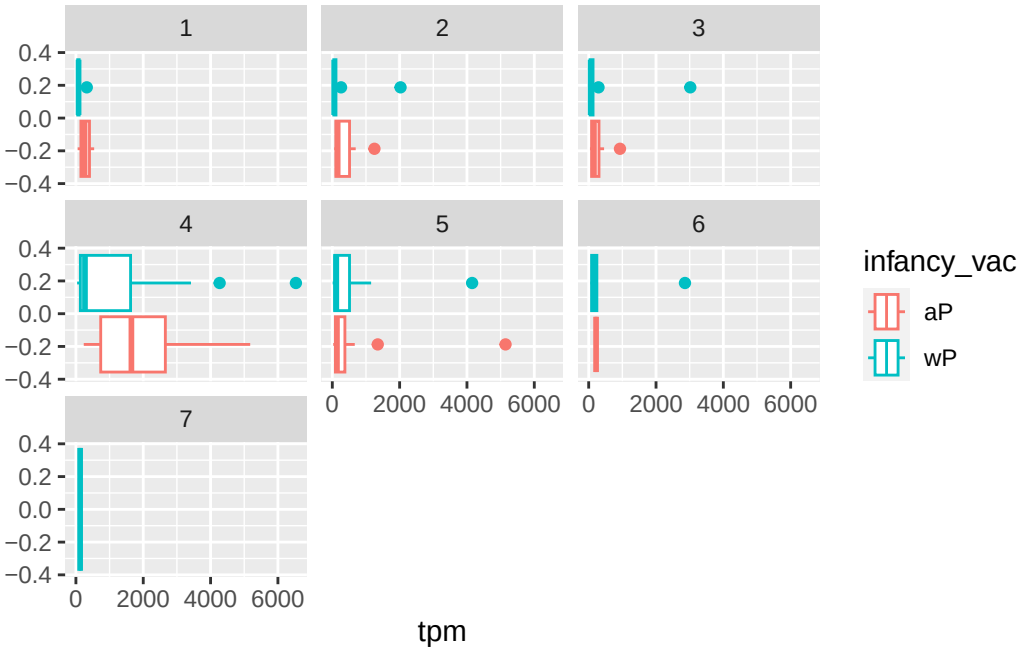
```
# Q19
ggplot(ssrna) +
  aes(visit, tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```



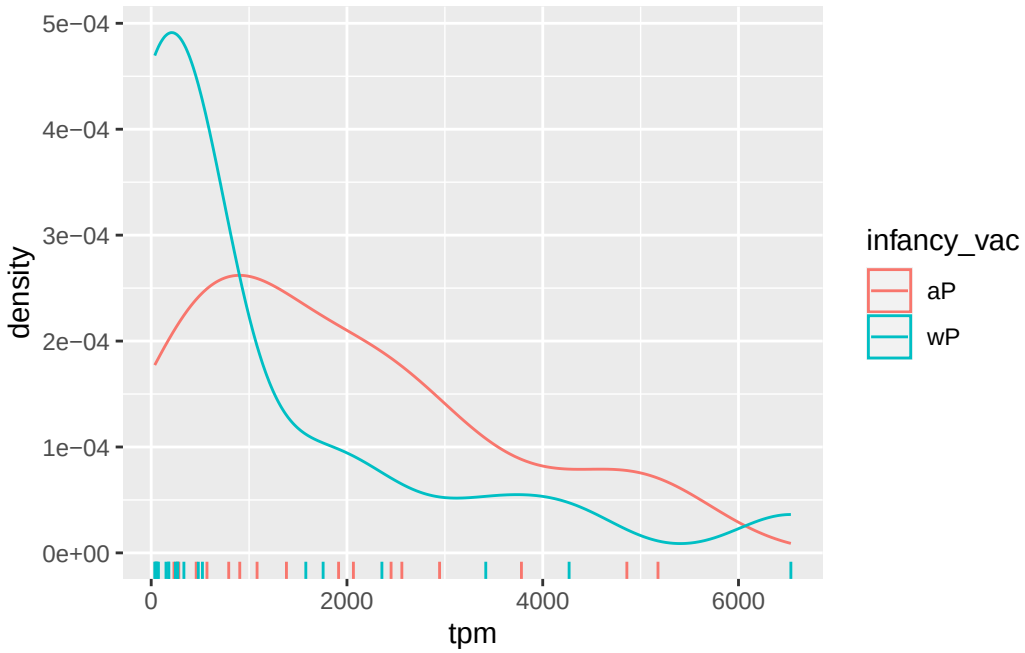
Q20. The expression level of this gene peaks at the 4th visit and drops after

Q21. It does not match the antibody production. Antibody level peaks at the 5th visit and is maintained at similar level until the 6th visit. Antibody needs more than to be translated into protein from mRNA, and since antibodies have relatively long half-lives, they can accumulate and are maintained in serum longer than mRNA.

```
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