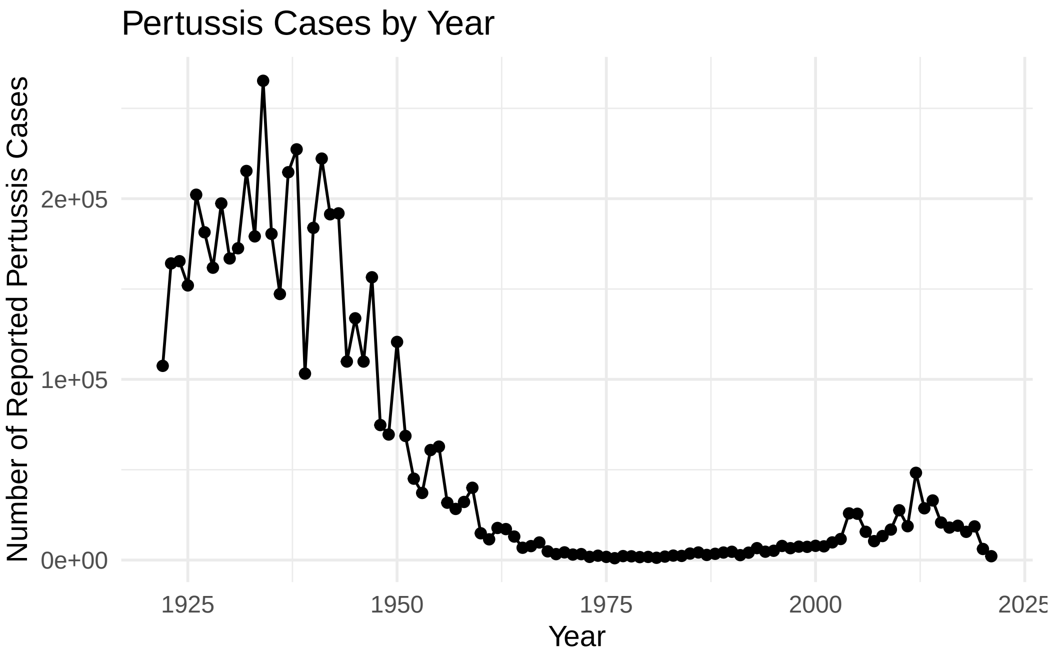
Class19

Alexander LIu (A69026918)

```
library(datapasta)
library(ggplot2)
```

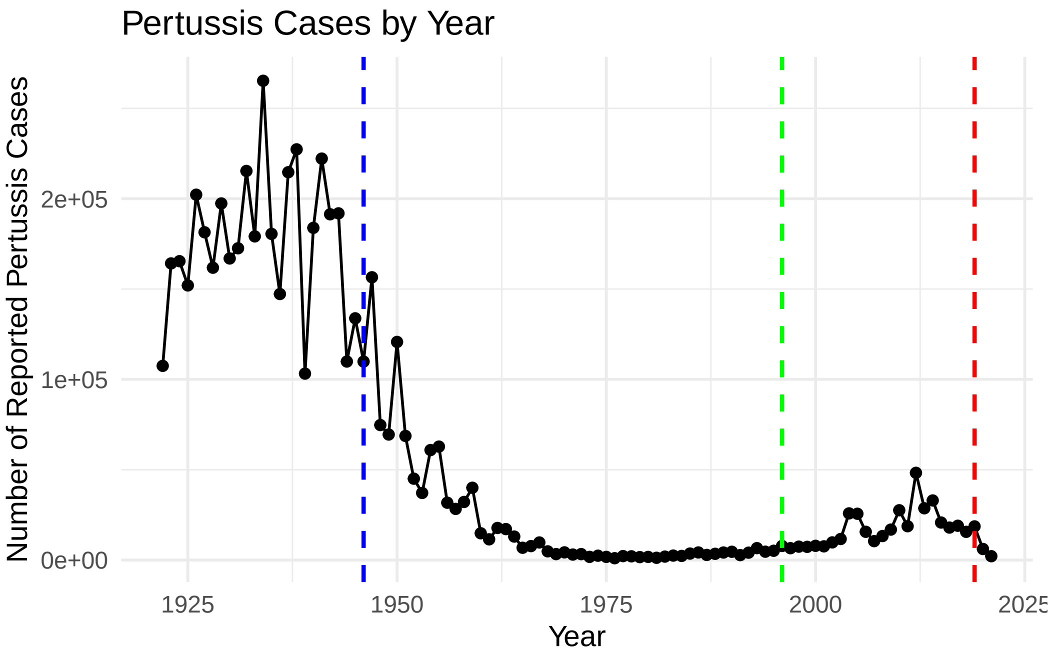
Q1.

```
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)
        )

g <- ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(title = "Pertussis Cases by Year", x = "Year", y = "Number of Reported Pertussis Ca
  theme_minimal()

g
```



```
g +
  geom_vline(xintercept = c(1946, 1996, 2019), linetype = "dashed", color = c("blue", "gre
```

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



Q2. The number of cases went down dramatically and rapidly after 1946. The number had been kept low since then, but after 2000 the number showed slight increase, until it started going down again in around 2010.

Q3. After 2000, the number showed slight increase, until it started going down again in around 2010.

```r
library(jsonlite)
```

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

```r
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

```
sum(subject$infancy_vac=="wP")
```

```
[1] 58
```

```
sum(subject$infancy_vac=="aP")
```

```
[1] 60
```

Q5

```
sum(subject$biological_sex=="Female")
```

```
[1] 79
```

```
sum(subject$biological_sex=="Male")
```

```
[1] 39
```

Q6

```
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
```

```r
library(lubridate)
```

```

Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```r
today()
```

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

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

```
Time difference of 8743 days
```

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

```
[1] 23.93703
```

Q7

```r
subject_1 <- subject
subject_1$age <- time_length(today() - ymd(subject_1$year_of_birth), "years")
```

```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
```

```
ap <- subject_1 %>% filter(infancy_vac == "aP")
round(summary(ap$age))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     21      26      26      26      27      30 
```

```
wp <- subject_1 %>% filter(infancy_vac == "wP")
round(summary(wp$age))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     28      31      35      36      39      56 
```

```
t.test(ap$age, wp$age)$p.value
```

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

(i) 36
(ii) 26
(iii) significantly different (p-value < 0.05)

Q8

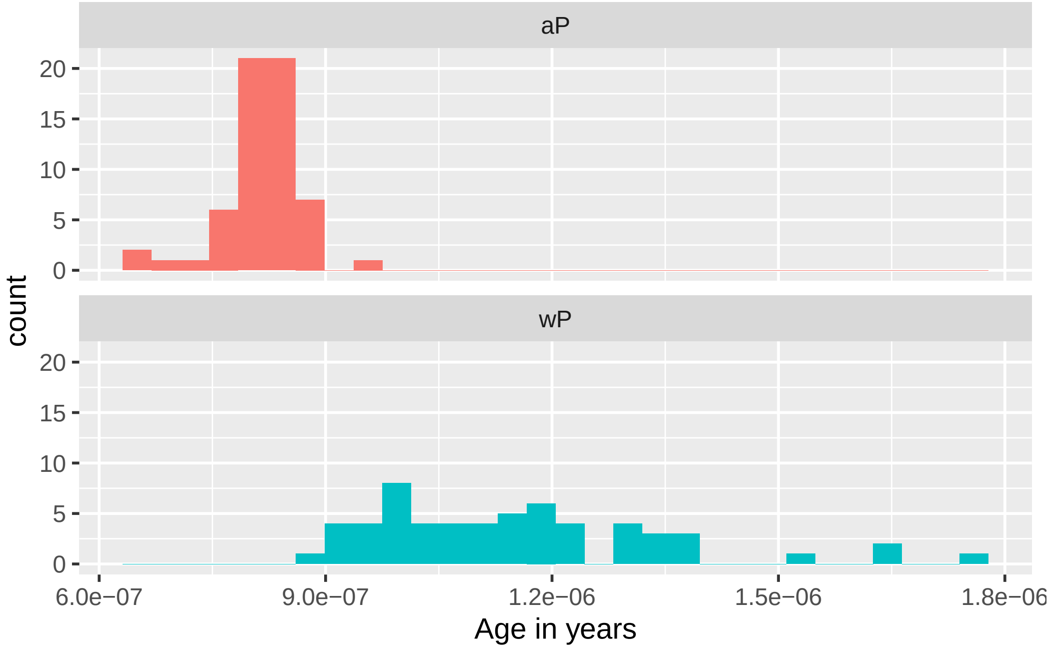
```
time_length( ymd(subject_1$date_of_boost) - ymd(subject_1$year_of_birth), "year")
```

```
  [1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481 35.84942 34.14921
  [9] 20.56400 34.56263 30.65845 34.56263 19.56194 23.61944 27.61944 29.56331
 [17] 36.69815 19.65777 22.73511 35.65777 33.65914 31.65777 25.73580 24.70089
 [25] 28.70089 33.73580 19.73443 34.73511 19.73443 28.73648 27.73443 19.81109
 [33] 26.77344 33.81246 25.77413 19.81109 18.85010 19.81109 31.81109 22.81177
 [41] 31.84942 19.84942 18.85010 18.85010 19.90691 18.85010 20.90897 19.04449
 [49] 20.04381 19.90691 19.90691 19.00616 19.00616 20.04381 20.04381 20.07940
 [57] 21.08145 20.07940 20.07940 20.07940 32.26557 25.90007 23.90144 25.90007
 [65] 28.91992 42.92129 47.07461 47.07461 29.07324 21.07324 21.07324 28.15058
 [73] 24.15058 24.15058 21.14990 21.14990 31.20876 26.20671 32.20808 27.20876
 [81] 26.20671 21.20739 20.26557 22.26420 19.32375 21.32238 19.32375 19.32375
 [89] 22.41752 20.41889 21.41821 19.47707 23.47707 20.47639 21.47570 19.47707
 [97] 35.90965 28.73648 22.68309 20.83231 18.83368 18.83368 27.68241 32.68172
[105] 27.68241 25.68378 23.68241 26.73785 32.73648 24.73648 25.79603 25.79603
[113] 25.79603 31.79466 19.83299 21.91102 27.90965 24.06297
```

Q9

```
ggplot(subject_1) +
  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`.
```



Significantly different.

Q9

```
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)
```

```
meta <- inner_join(specimen, subject_1)
```

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

```r
dim(meta)
```

```
[1] 939  14
```

Q10

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

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

```r
dim(abdata)
```

```
[1] 41810    21
```

Q11

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

```

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

Q12

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

```

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

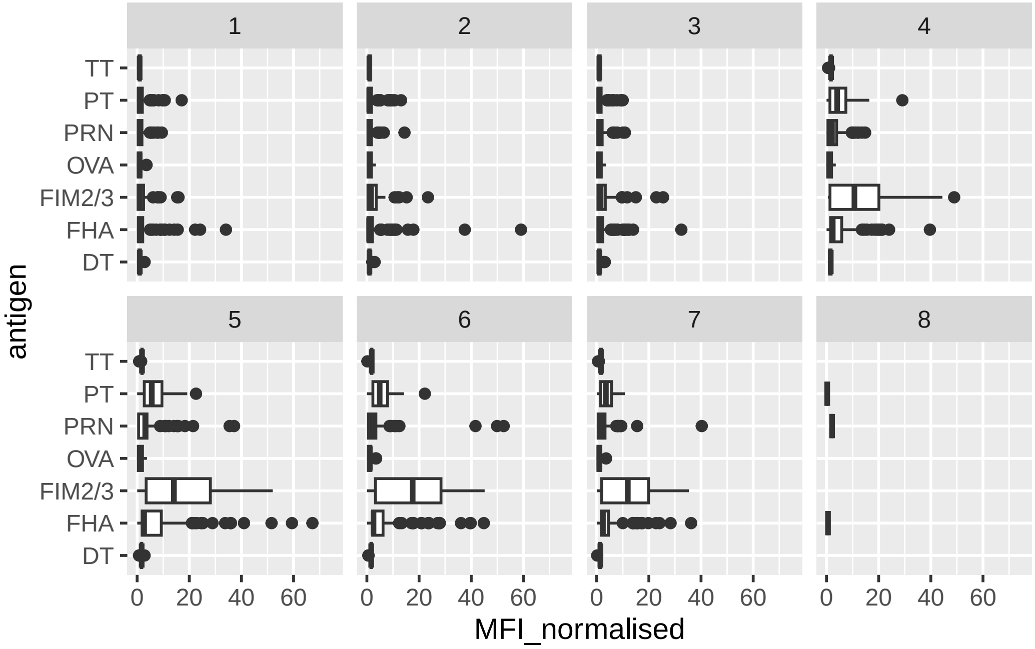
```r
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

```
  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
1 37.93566
2 37.93566
3 37.93566
4 40.93634
5 40.93634
6 40.93634
```

Q13

```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()`).
```

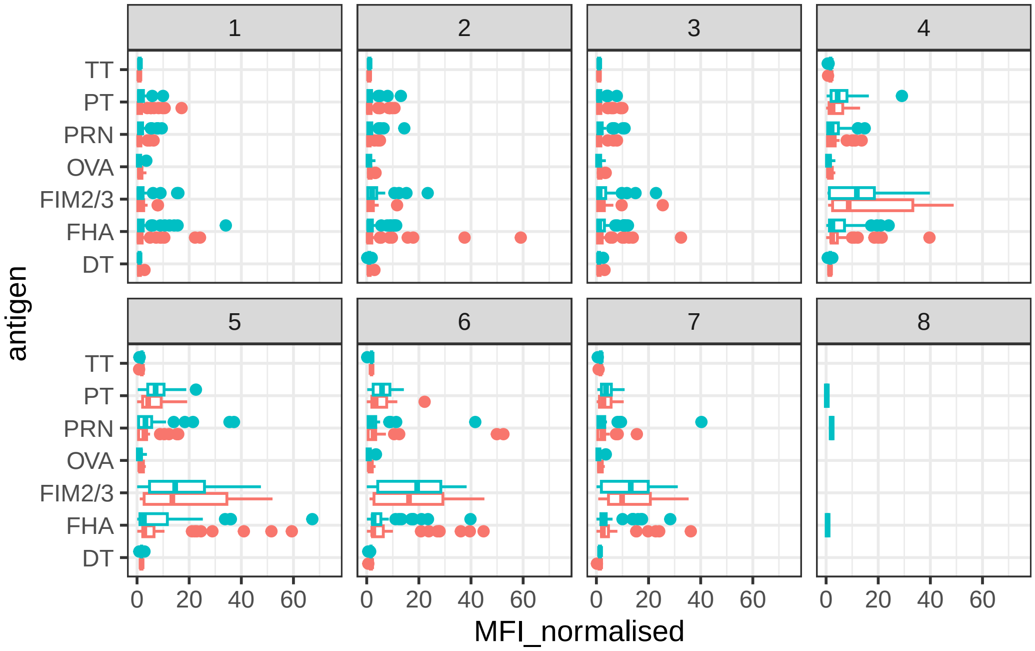


Q14

IgG that are against PT, PRN, FIM2/3 and FHA showed differences in the level, because these are includede in the vaccine. As opposed to there, TT, OVA, and DT are not included in the vaccine, so IgG against them were not induced.

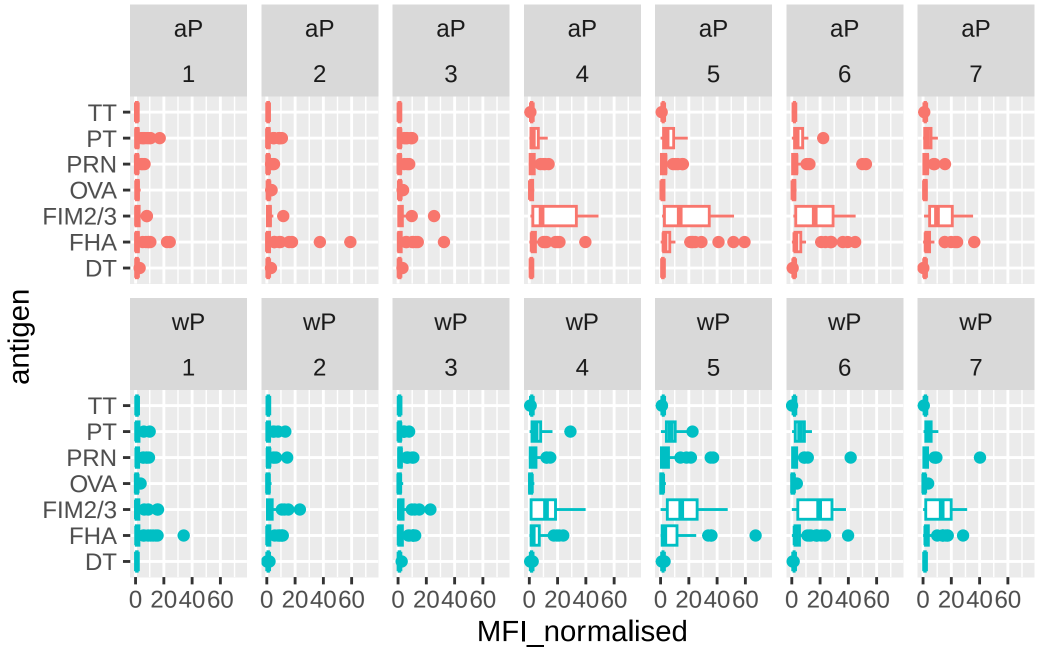
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

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



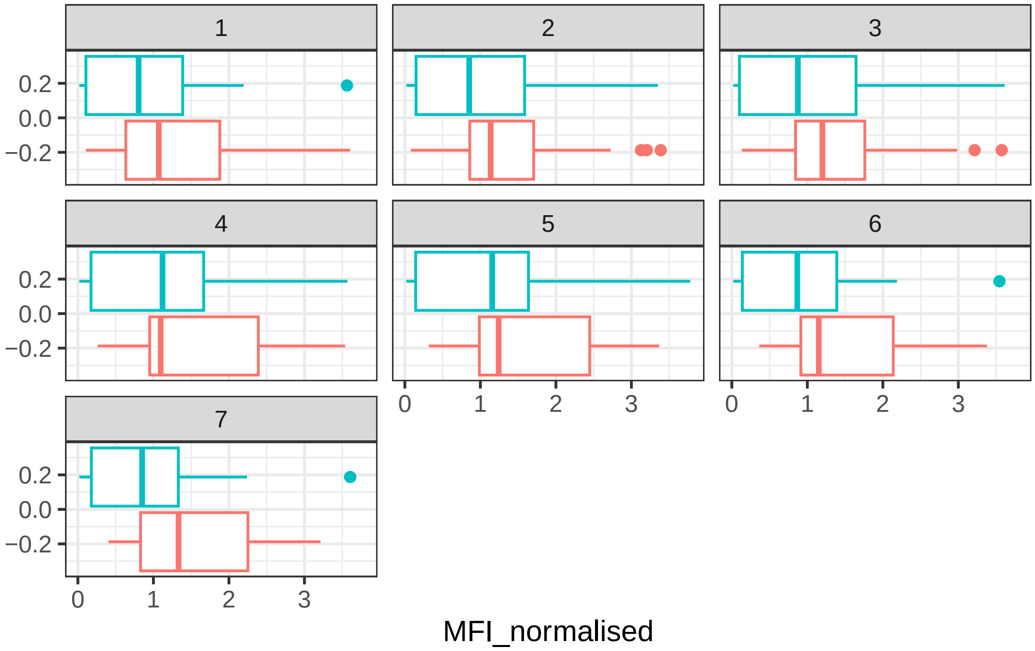
```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

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

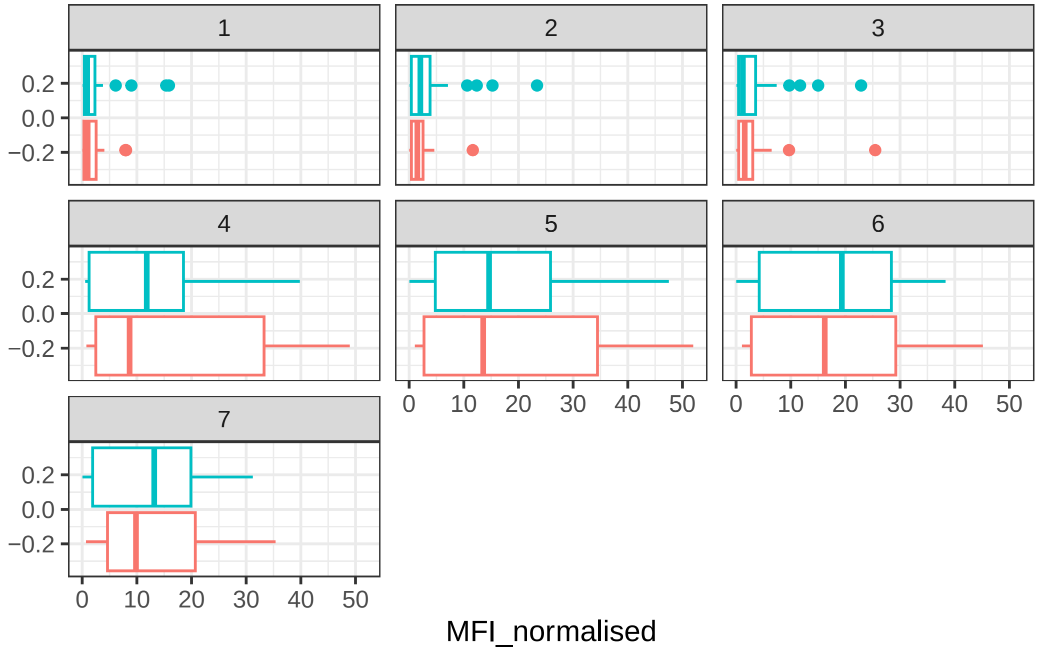


Q15

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



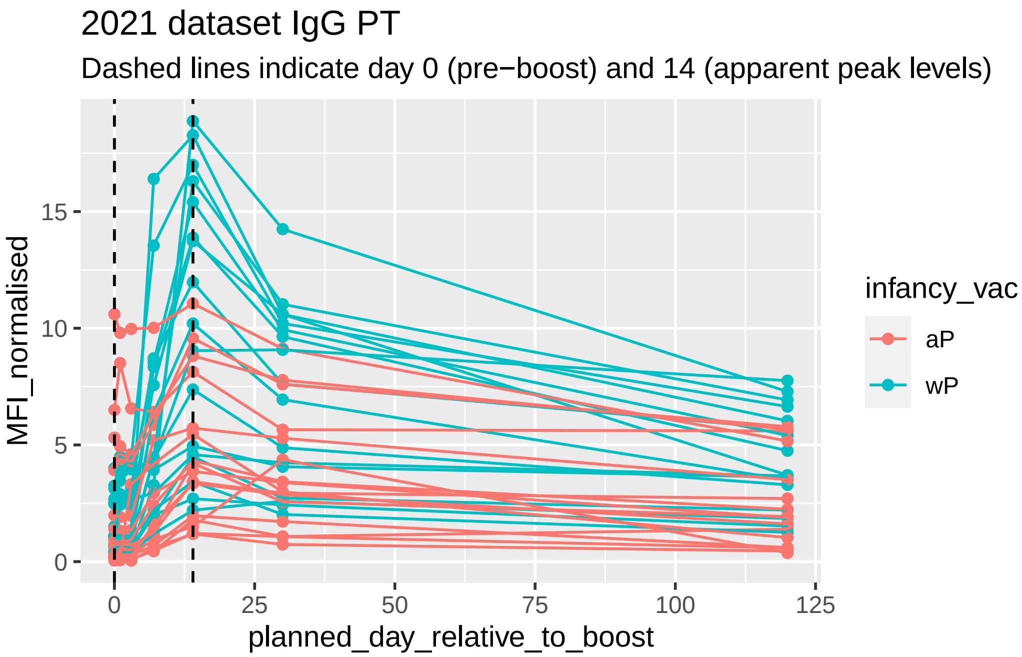
```
filter(igg, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Q16 The level of anti-PT IgG increases over time and decreases after peaking at visit 5, while anti-OVA stays about the same throughout the visits. Also, the level of anti-PT is much higher than anti-OVA IgG.

Q17 No. They are overall similar.

```
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)")
```



```
wP_abdata.21 <- abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT", infancy_vac == "wP")
aP_abdata.21 <- abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT", infancy_vac == "aP")

t.test(wP_abdata.21$MFI_normalised, aP_abdata.21$MFI_normalised)$p.value
```

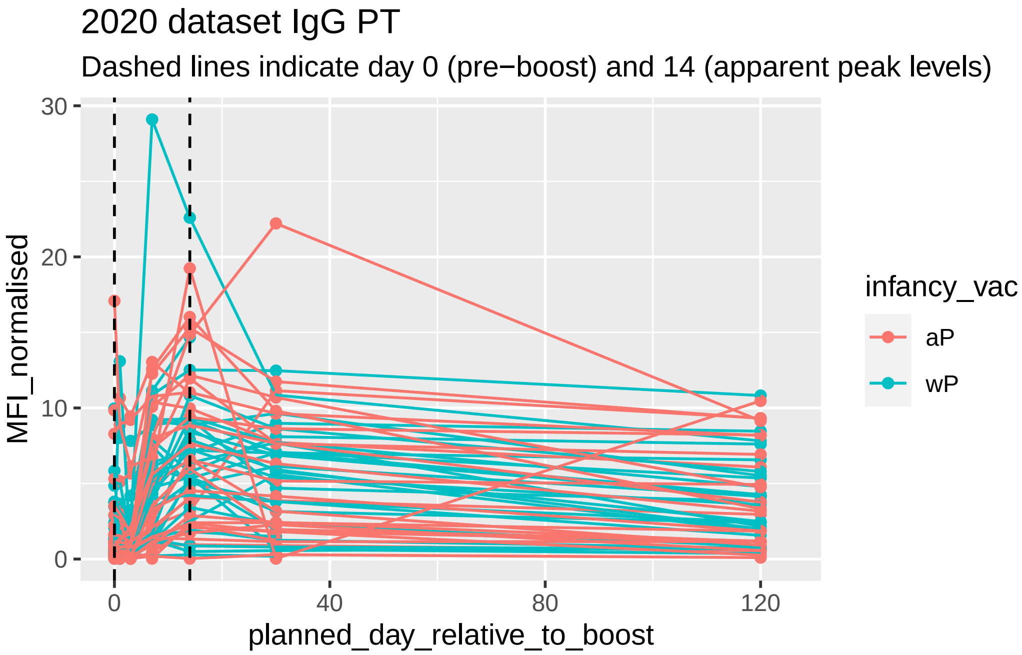
```
[1] 0.0003848114
```

Q18

```
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()`).
```



```
wP_abdata.20 <- abdata.20 %>%
  filter(isotype == "IgG",  antigen == "PT", infancy_vac == "wP")
aP_abdata.20 <- abdata.20 %>%
  filter(isotype == "IgG",  antigen == "PT", infancy_vac == "aP")

t.test(wP_abdata.20$MFI_normalised, aP_abdata.20$MFI_normalised)$p.value
```

```
[1] 0.4907405
```

In 2021 anti-PT IgG level has overall higher levels, while in 2020 it has more similar levels. This is confirmed by p-values from t-test, 0.0003848114 and 0.4907405, respectively for 2021 and 2020.

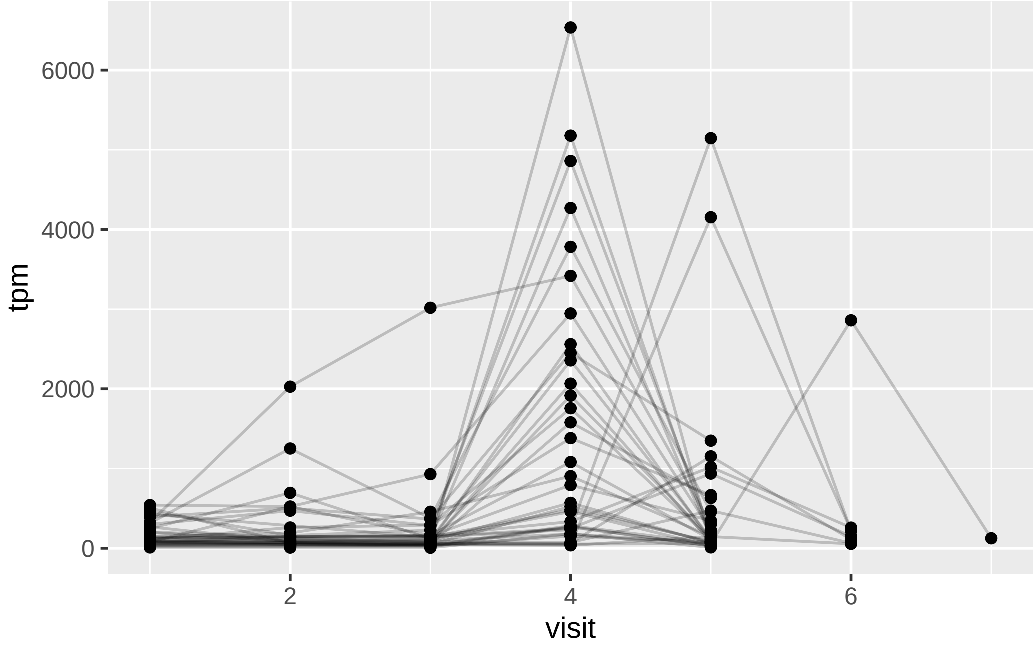
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.

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

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

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

Q19

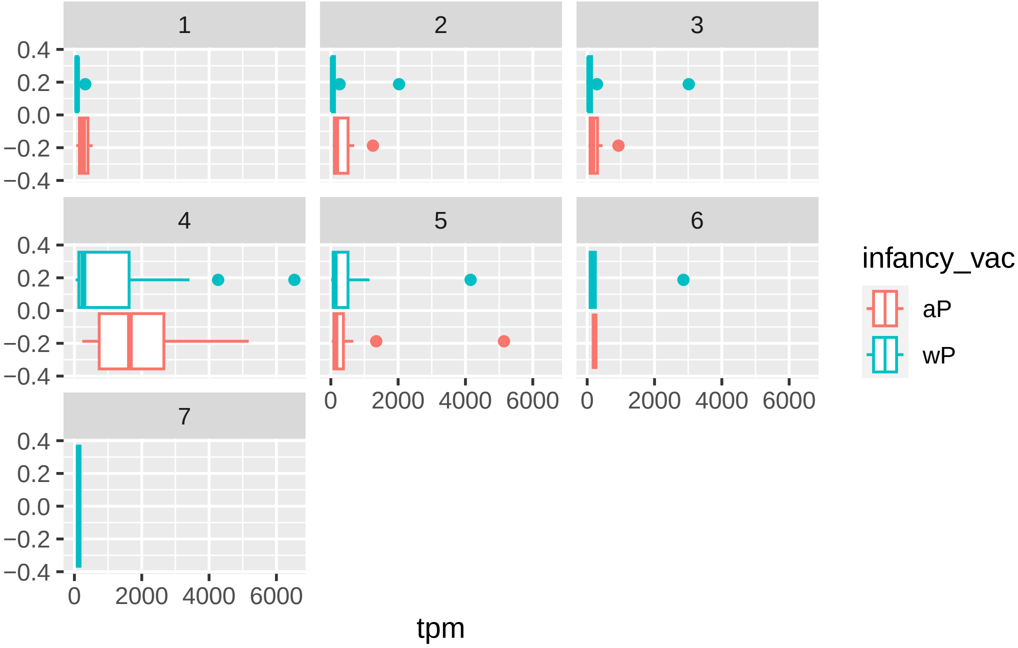
```
ggplot(ssrna) +
  aes(visit, tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```



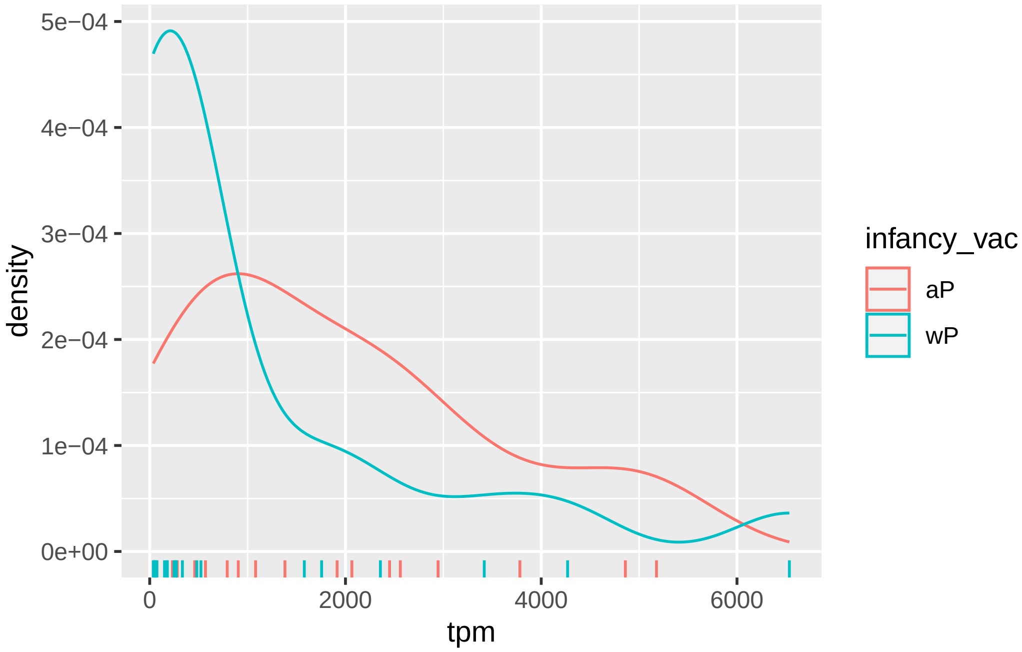
Q20 Visit 4.

Q21 They do not match. Transcripts and proteins are different in half-lives, and the time for production.

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