Lab 15: Mini Project: Investigating Pertussis Resurgence

Pamelina Lo (AID: 16735368)

Background

Pertussis, aka Wooping Cough, is a highly infectious lunch disease caused by the bacteria *B. Pertussis*.

The CDC tracks pertussis cases numbers per year. Lets have a look at this data:CDC data

We will use the **datapasta** R package to "scrape" this data into R.

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.

```
options(repos = c(CRAN = "https://cran.rstudio.com/"))
install.packages("datapasta")
```

```
The downloaded binary packages are in
    /var/folders/bc/dmxqsptj30x2fv4dj93n0fw00000gn/T//RtmpmFk5fa/downloaded_packages
```

```
library(datapasta)
```

```
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, 2024L),
        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, 23544)
       )
```

```
library(ggplot2)
baseplot <- ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(title = "Petussiss Cases by Year", y = "Number of Cases")
```

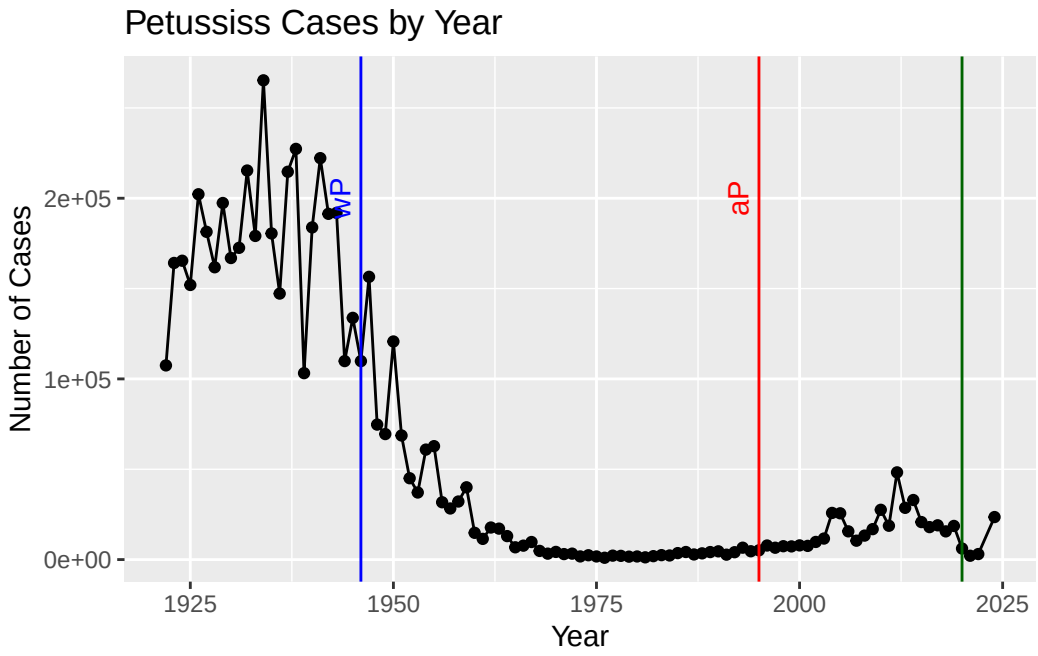
##2. A tale of two vaccines (wP & aP)

Add some landmarks developments as annoations to out plot. We include the first whole-cell (wP) caccine roll-out in 1946.

Let's add the switch to acellular vaccine (aP) in 1996.

> 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?

```
baseplot +
  geom_vline(xintercept = 1946, col="blue") +
  annotate("text", x = 1946, y = 200000, label = "wP", col = "blue", angle = 90, vjust = -0.5
  geom_vline(xintercept = 1995, col = "red") +
 annotate("text", x = 1995, y = 200000, label = "aP", col = "red", angle = 90, vjust = -0.5)
  geom_vline(xintercept = 2020, col="darkgreen")
```



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

From the beginning, the cases went to ~200,000 cases pre wP vaccine to ~1,000 cases in 1976. The US switched to the aP vaccine in 1995 which we see a big increase in 2004 to ~26,000 cases. The possible explanations for this trend is bacterial evolution due to COVID 19 pandemic or the vaccination rates are getting lower. Additionally, humans gain stronger immune response over time in which aP vaccine must be re-administered help to build immunity. Unfortunately, people don't get re-vaccinated, resulting more cases of individuals infected.

There is a ~ 10 year long lag from aP roll out to increasing cases numbers. THis hold true of other coutnreis like Japan, UK, etc.

Key Question: Why does the aP vaccine induced immunity wane faster than that of the wP vaccine? The aP vaccine induced immunity wane faster than the wP vaccine because the aP vaccine stimulate a dominate immune response which produces an abundant of antibodies.

##3. Exploring CMI-PB

The CMI-PB (Computational Models of Immunity Pertussis Boost) makes available lots of data about the immune response to Pertussis vaccination.

Crtitically, it tracks wP and aP individuals over time to see how their immune response changes.

CMI-PB make all their data freely available via JSON format tables from their database.

Lets read the first one of these tables:

```
library(jsonlite)
subject <- read_json("http://cmi-pb.org/api/v5/subject" ,
                     simplifyVector = TRUE)
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
```

Class Question: How many individuals are in this data?

```
nrow(subject)
```

```
[1] 172
```

There are 172 individuals.

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

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

```
aP wP
87 85
```

There are 87 aP and 85 wP individuals.

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

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

```
Female   Male
   112     60
```

There are 112 females and 60 males.

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

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

```
                                        Female Male
  American Indian/Alaska Native              0    1
  Asian                                     32   12
  Black or African American                  2    3
  More Than One Race                        15    4
  Native Hawaiian or Other Pacific Islander  1    1
  Unknown or Not Reported                   14    7
  White                                     48   32
```

Class Question: Does this do a good job representing the US populus?

No, this is not a good representation of the US populus because we are limited with much individuals.

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2024-11-19"
```

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

```
Time difference of 9089 days
```

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

```
[1] 24.88433
```

> 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?

(i) the average age of wP individuals

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

```
[1] 38.88296 56.88433 41.88364 36.88433 33.88364 36.88433
```

```
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
```

```
# wP
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
```

The average wP individuals are 36 individuals.

(ii) the average age of aP individuals

```
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
```

The average aP individuals are 27 individuals.

(iii) are they significantly different They are significantly different, but lets check with a p-test. If the p-test is lower than 0.05, then they are significantly different.

```
t_test_result <- t.test(wp$age, ap$age)
print(t_test_result)
```

```
    Welch Two Sample t-test

data:  wp$age and ap$age
t = 12.918 days, df = 104.03, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2705.535 days 3686.855 days
sample estimates:
Time differences in days
mean of x mean of y
12977.471  9781.276
```

Since the p-value is smaller than 0.05, then aP and wP are significantly different.

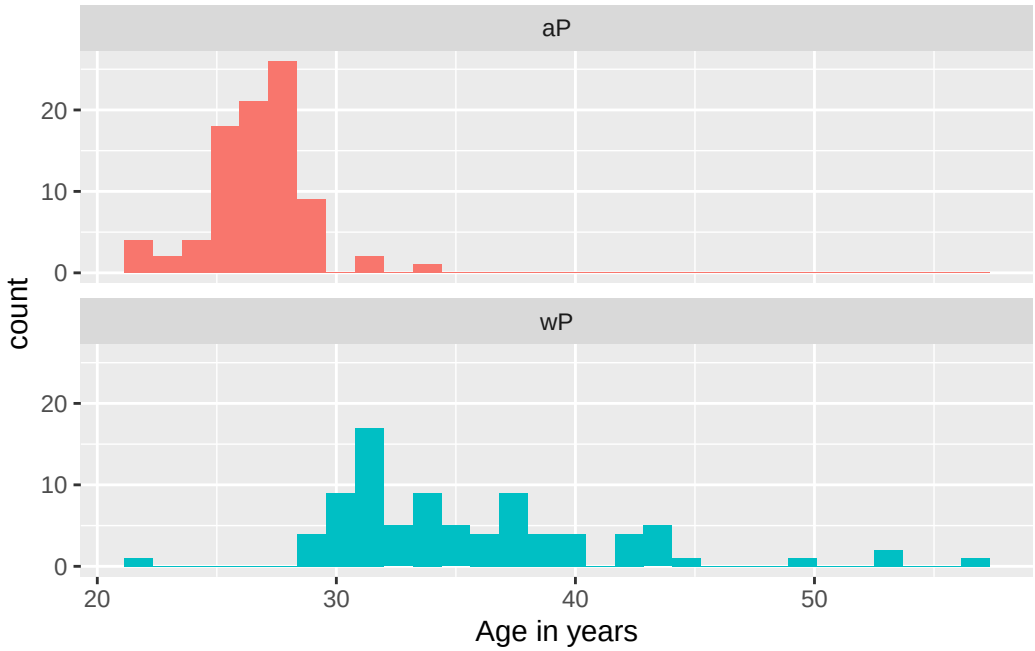
Q8. Determine the age of all individuals at time of boost?

```
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(x = time_length(age, "year"), fill = as.factor(infancy_vac)) +
  geom_histogram(bins = 30, show.legend = FALSE) +
  facet_wrap(vars(infancy_vac), nrow = 2) +
  xlab("Age in years")
```




```
x <- t.test(time_length( wp$age, "years" ),
       time_length( ap$age, "years" ))

x$p.value
```

```
[1] 2.372101e-23
```

I think these two groups are significantly different because, from the t-test, the p-value is smaller than 0.05.

##Joining multiple tables

Lets get more data from CMI-PB, this time about the specimens collected.

```
specimen <- read_json("http://cmi-pb.org/api/v5/specimen",
                      simplifyVector = TRUE)
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
```

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:

Now we can join/merge these two tables `subject` and `specimen` to make one new `meta` table table with the combined data.

```
library(dplyr)

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

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

```
head(meta)
```

```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
3          1          wP         Female Not Hispanic or Latino White
4          1          wP         Female Not Hispanic or Latino White
5          1          wP         Female Not Hispanic or Latino White
6          1          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset        age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 14202 days           1
2    1986-01-01    2016-09-12 2020_dataset 14202 days           2
3    1986-01-01    2016-09-12 2020_dataset 14202 days           3
4    1986-01-01    2016-09-12 2020_dataset 14202 days           4
5    1986-01-01    2016-09-12 2020_dataset 14202 days           5
6    1986-01-01    2016-09-12 2020_dataset 14202 days           6
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                           -3                             0         Blood
```

```
2                              1                              1         Blood
3                              3                              3         Blood
4                              7                              7         Blood
5                             11                             14         Blood
6                             32                             30         Blood
  visit
1     1
2     2
3     3
4     4
5     5
6     6
```

Now read an "experiment data" table from CMI-PB

```
abdata <- read_json("http://cmi-pb.org/api/v5/plasma_ab_titer",
                    simplifyVector = TRUE)
head(abdata)
```

```
  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
```

One more joint to do of `meta` anda `abdata` to associate all the metadata about the individual and their race, biological sex and infanticy vaccination status together with Antibody levels.

> 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.

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

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

```
head(ab)
```

```
  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 subject_id infancy_vac biological_sex
1 UG/ML                 2.096133          1          wP         Female
2 IU/ML                29.170000          1          wP         Female
3 IU/ML                 0.530000          1          wP         Female
4 IU/ML                 6.205949          1          wP         Female
5 IU/ML                 4.679535          1          wP         Female
6 IU/ML                 2.816431          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 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 actual_day_relative_to_boost planned_day_relative_to_boost
1 14202 days                          -3                             0
2 14202 days                          -3                             0
3 14202 days                          -3                             0
4 14202 days                          -3                             0
5 14202 days                          -3                             0
6 14202 days                          -3                             0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

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

```
table(ab$isotype)
```

```
  IgE   IgG  IgG1  IgG2  IgG3  IgG4
 6698  5389 10117 10124 10124 10124
```

How many antigens?

```
table(ab$antigen)
```

```
   ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA
  1970    1970    4978    1970    5372    4978    1970    1970    1970    4978
   PD1     PRN      PT     PTM   Total      TT
  1970    5372    5372    1970     788    4978
```

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(ab$dataset)
```

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

From this dataset, I've noticed that the number individuals of the 2023 dataset is smaller. This suggests that from 2020-2023 there has been a decrease number of individuals who were tested for Pertussis.

##4. Examine IgG Ab titer levels Lets focus on IgG - one of the main antibody types responsive to bacteria or viral infections.

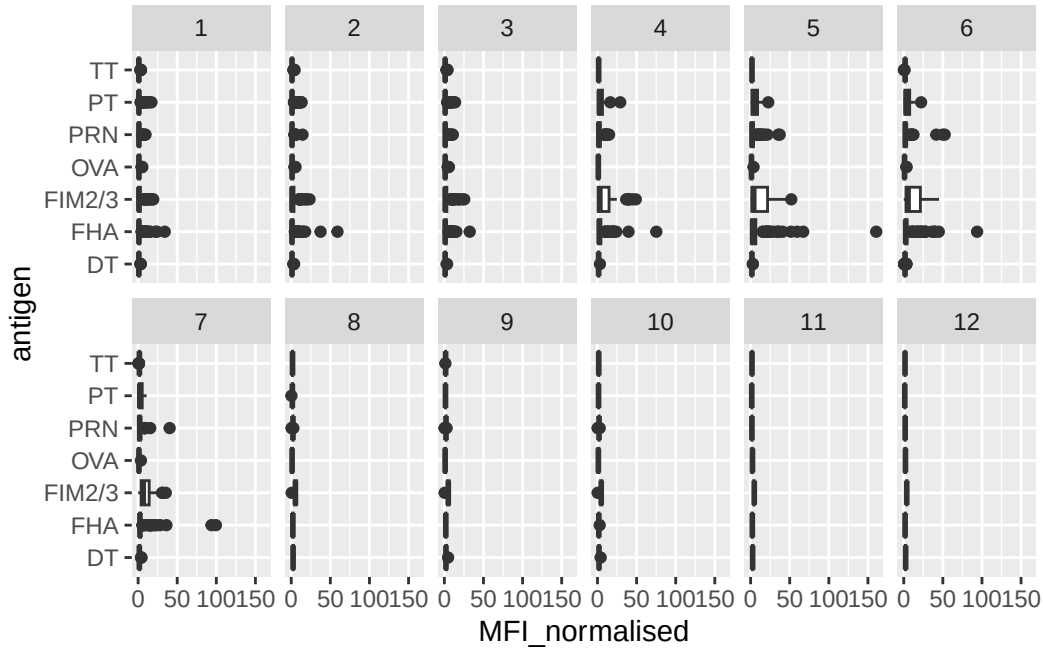
```
igg <- filter(ab, isotype=="IgG")
head(igg)
```

```
  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 infancy_vac biological_sex
1 IU/ML                 0.530000          1          wP         Female
2 IU/ML                 6.205949          1          wP         Female
3 IU/ML                 4.679535          1          wP         Female
4 IU/ML                 0.530000          3          wP         Female
5 IU/ML                 6.205949          3          wP         Female
6 IU/ML                 4.679535          3          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 actual_day_relative_to_boost planned_day_relative_to_boost
1 14202 days                          -3                             0
2 14202 days                          -3                             0
3 14202 days                          -3                             0
4 15298 days                          -3                             0
5 15298 days                          -3                             0
6 15298 days                          -3                             0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

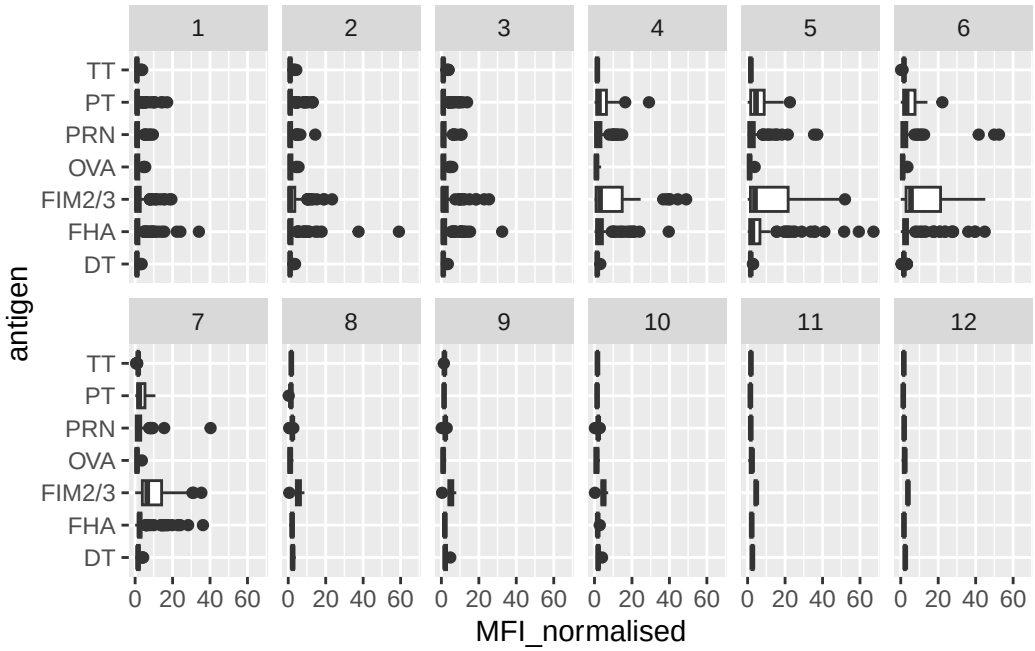
Make a fitst plot of MFI (Meam Flurosence Intensity - measure of how much is detected) for each antigen.

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(0,161) +
  facet_wrap(vars(visit), nrow=2)
```




```
ggplot(igg %>% filter(MFI_normalised >= 0 & MFI_normalised <= 75)) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow = 2)
```

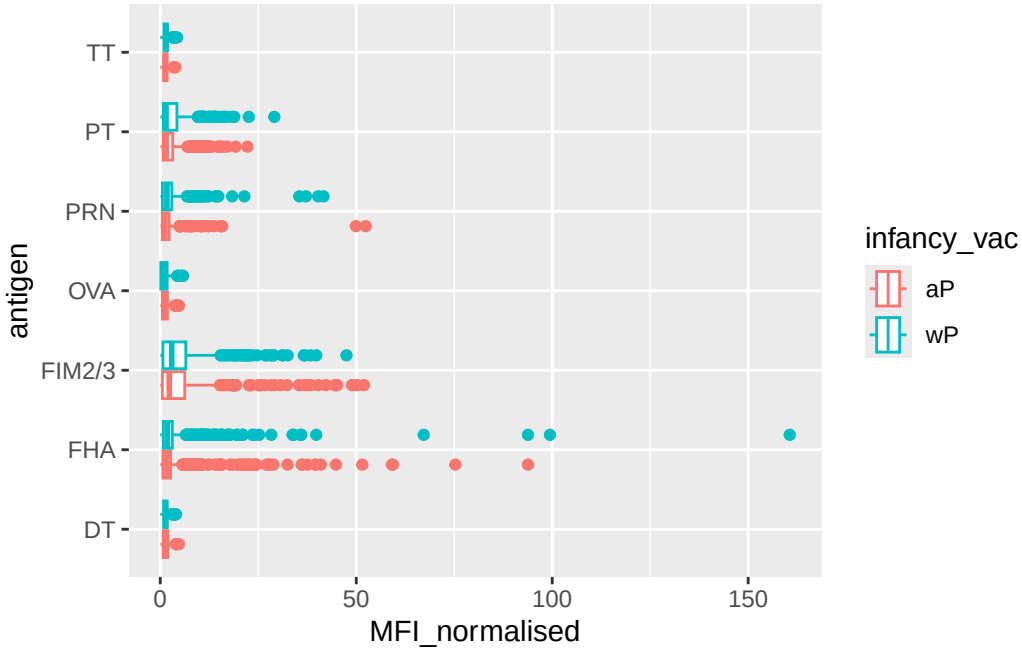


Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

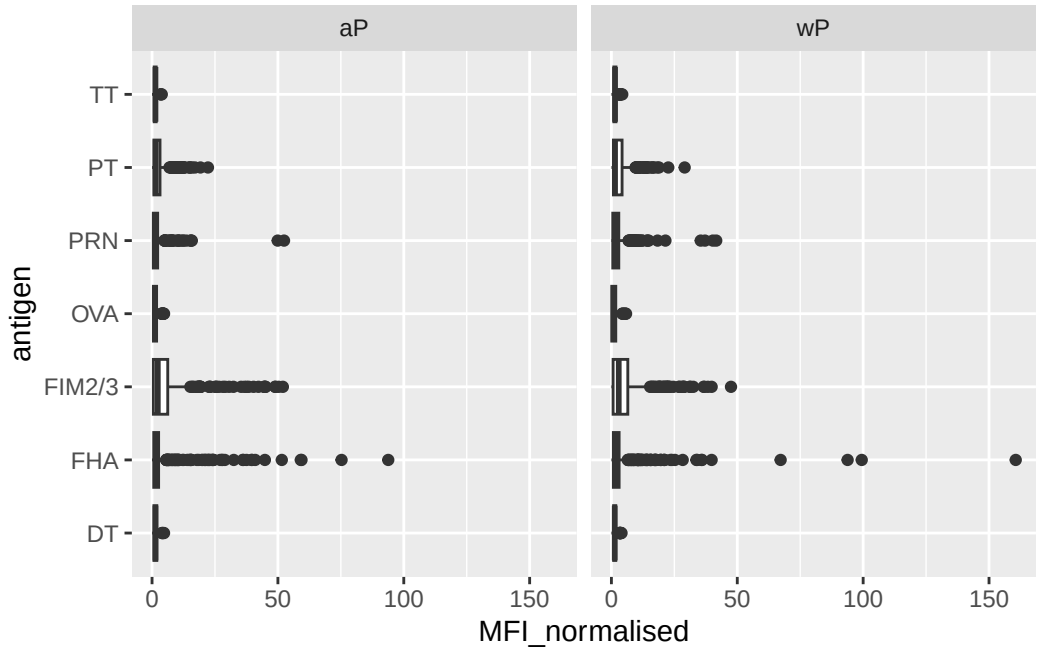
PT, PRN, FIM2/3, and FHA antigens show difference in the level of IgG antibody titers. These antigens show these differences because their functionality is more accessible for antibody binding and immune activation, for instance PT functionality is to be accessibly to the immune system which makes it capable to generate antibody responses. PRN and FHA is located on the bacterial surface which assists with binding to antibodies for immune activation.

Examine differences between wP and aP We can attempt to examine differences between wP and aP here by setting color and/or facet values of the plot to include infancy_vac status (see below). However these plots tend to be rather busy and thus hard to interpret easily.

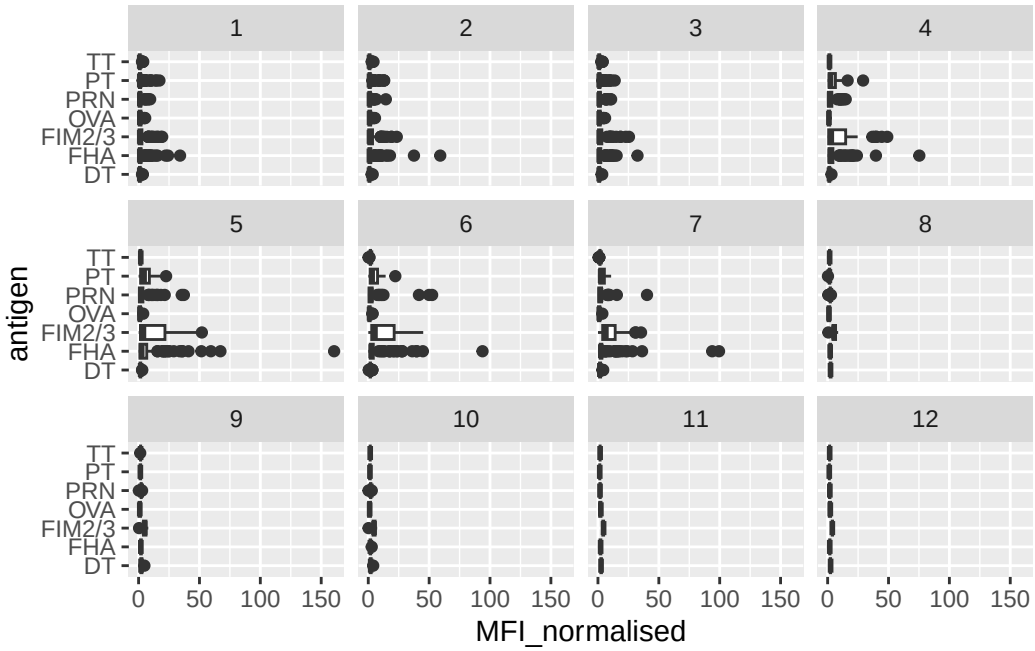
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac) +
  geom_boxplot()
```




```
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
  facet_wrap(~infancy_vac)
```

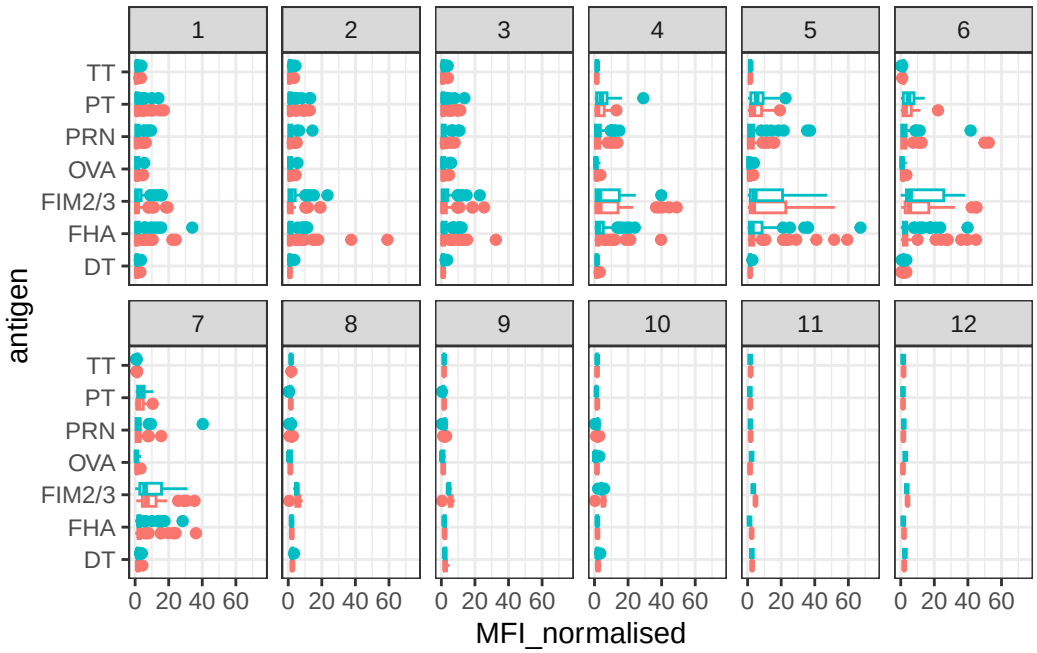



```
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
  facet_wrap(~visit)
```



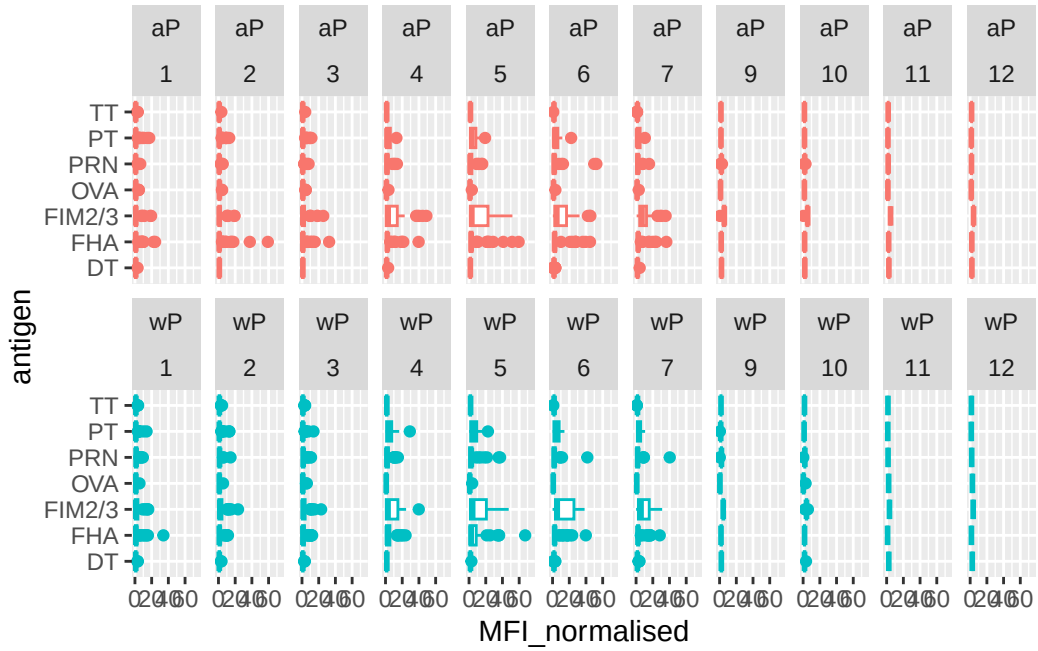

```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

```
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()`).
```




```
table(igg$visit)
```

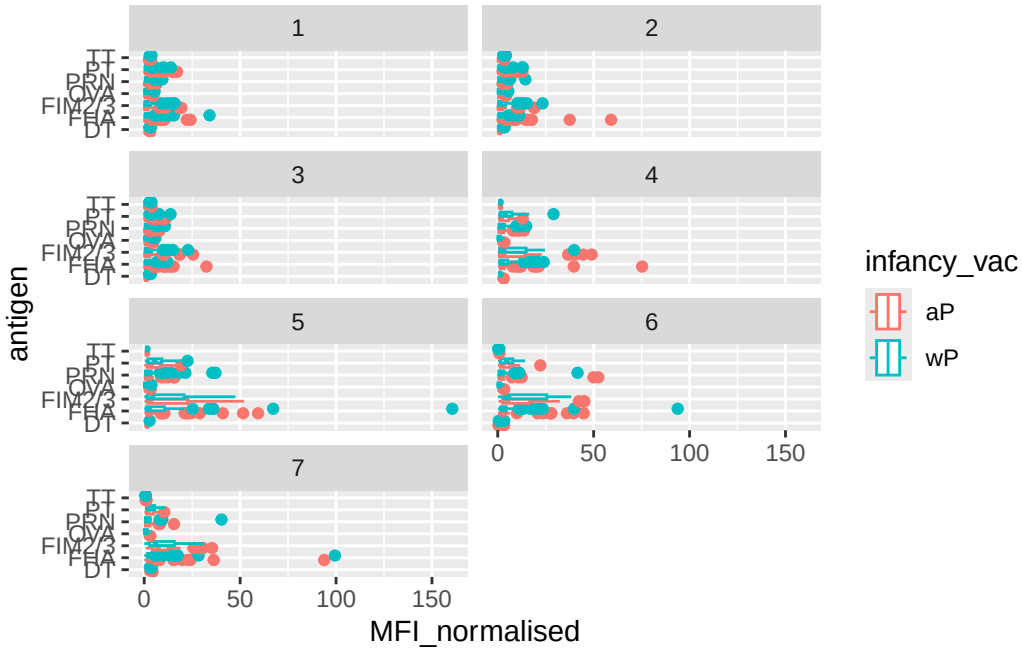
```
  1   2   3   4   5   6   7   8   9  10  11  12
902 902 930 559 559 540 525 150 147 133  21  21
```

Looks like we don't have data yet for all subjects in terms of visits 8 onwards. So lets exclude these.

```
igg_7 <- filter(igg, visit %in% 1:7)
table(igg_7$visit)
```

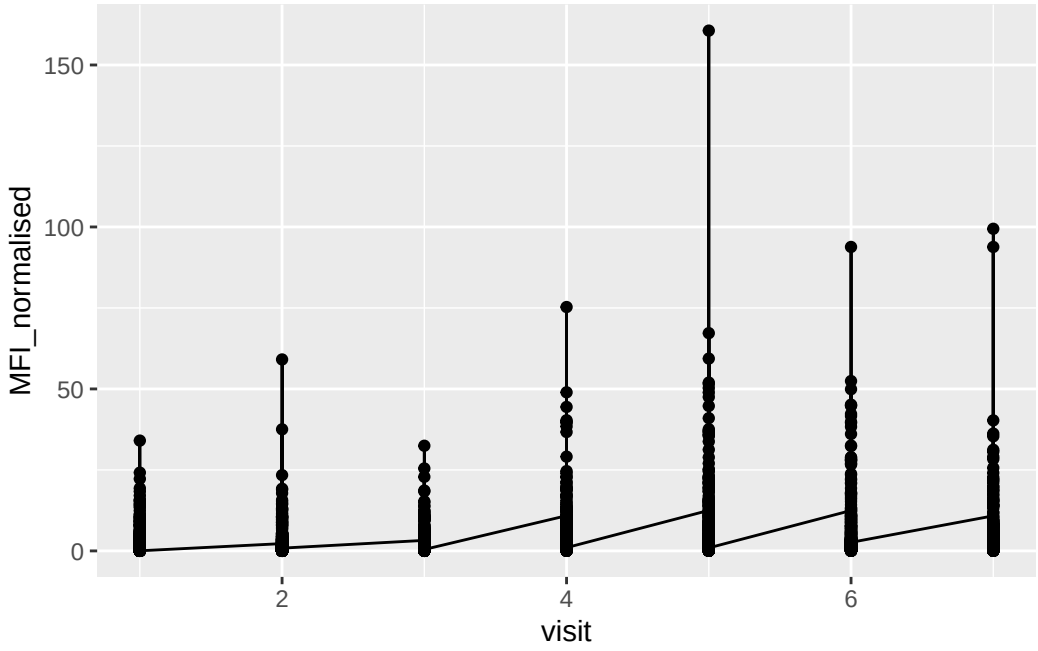
```
  1   2   3   4   5   6   7
902 902 930 559 559 540 525
```

```
ggplot(igg_7) +
  aes(MFI_normalised, antigen, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(~visit, ncol=2)
```



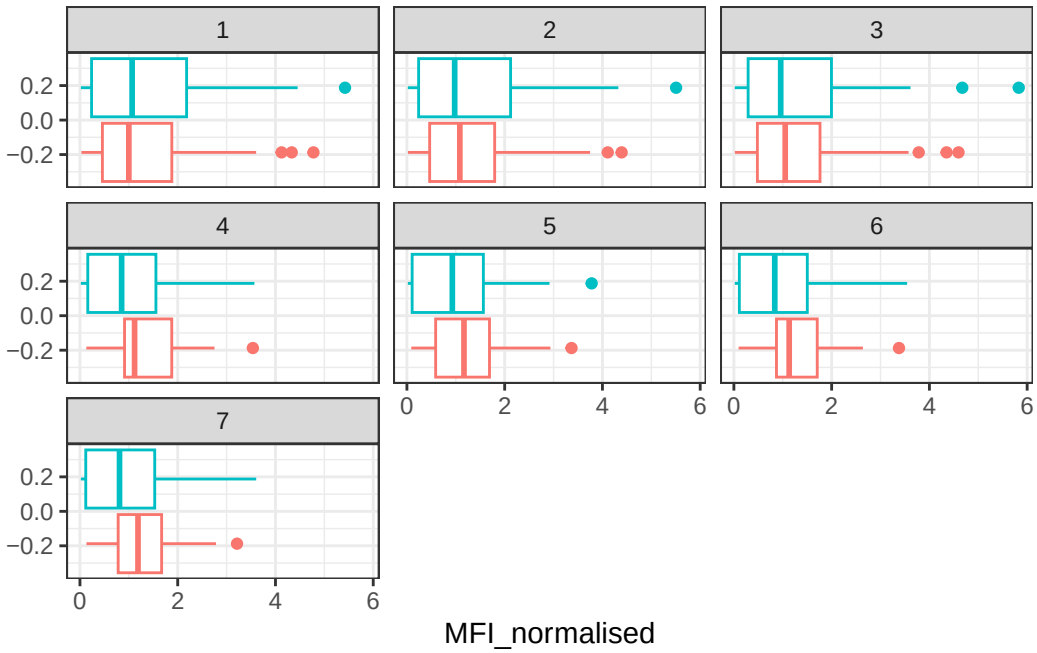
Let's try a different plot. First focus on one antigen, start with PT (Pertussiss Toxin) and plot visits or time on the x-axis and the MFI_normalized on the y-axis.

```
ggplot(igg_7) +
  aes(visit, MFI_normalised, group) +
  geom_point() +
  geom_line()
```



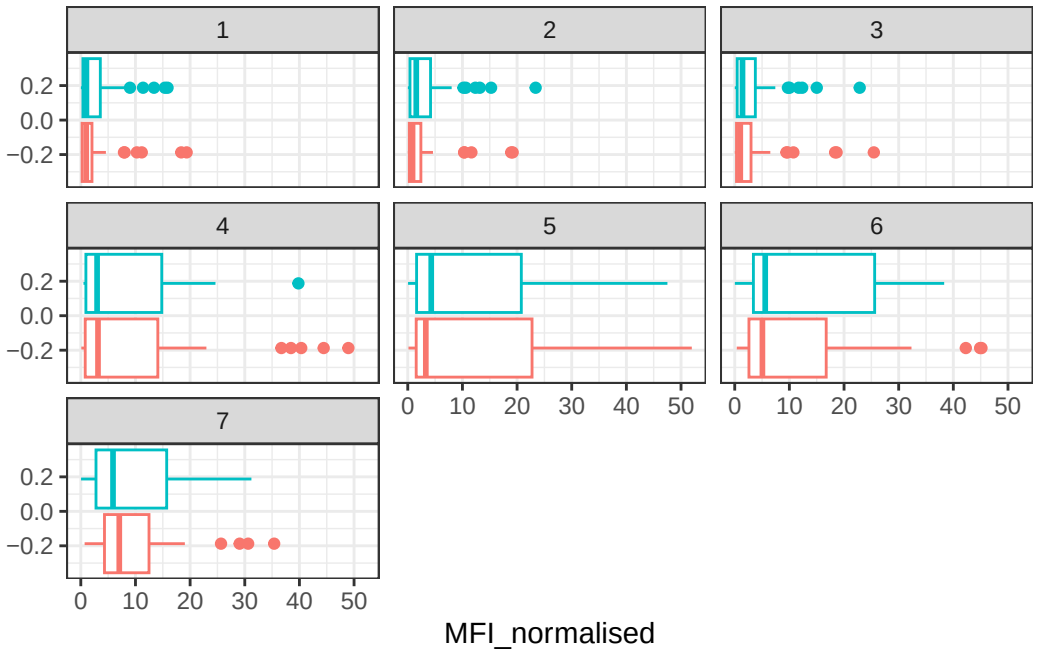
Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium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

```
filter(igg_7, antigen=="OVA") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



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

```
filter(igg_7, antigen=="FIM2/3") %>%
  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?

I've noticed that Pertussis Toxin levels have risen over time on both antigens. However, there rise that appears more apparent the OVA antigen boxplot than the FIM2/3 antigen boxplot. Both boxplots show cases that the patients coming for their fifth visit is where the PT data peaks and declines. This trend appears in both wP and aP subjects.

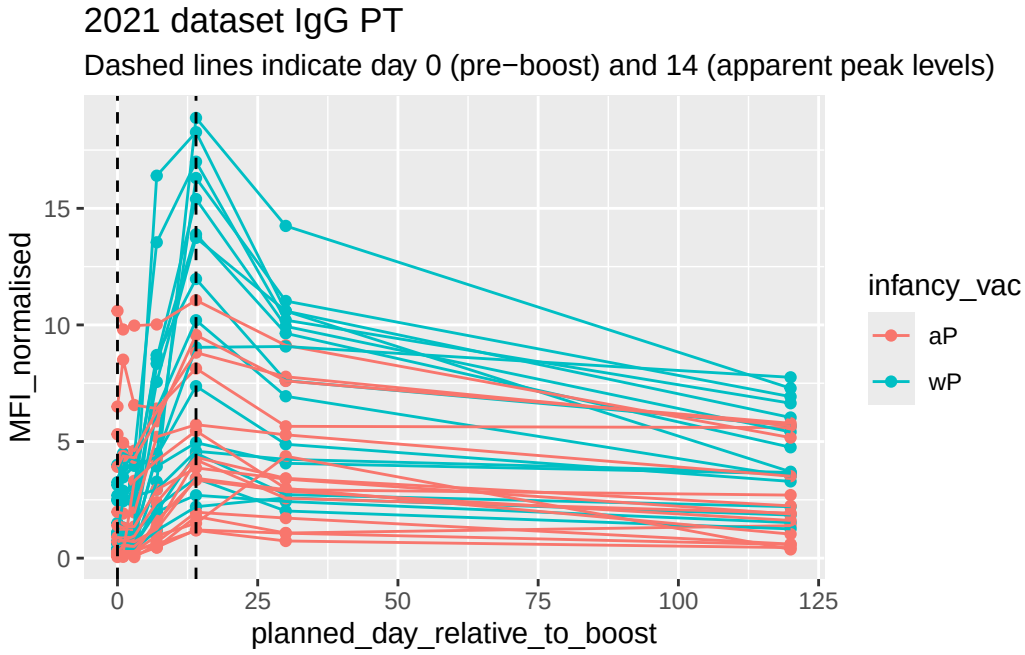
Q17. Do you see any clear difference in aP vs. wP responses?

There is a difference in aP vs wP responses. In the OVA antigen boxplot, the data shows that aP patients have higher levels of PT levels compared to wP responses. In comparision to the FIM2/3 antigen boxplot, the data shows that wP responses have slightly higher levels of PT levels. This suggests that aP patients with the OVA antigen will experience higher PT levels where as wP patients with the FIM2/3 antigen will experience higher PT levels.

Lets finish this section by looking at the 2021 dataset IgG PT antigen levels time-course:

```
abdata.21 <- ab %>% 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)")
```

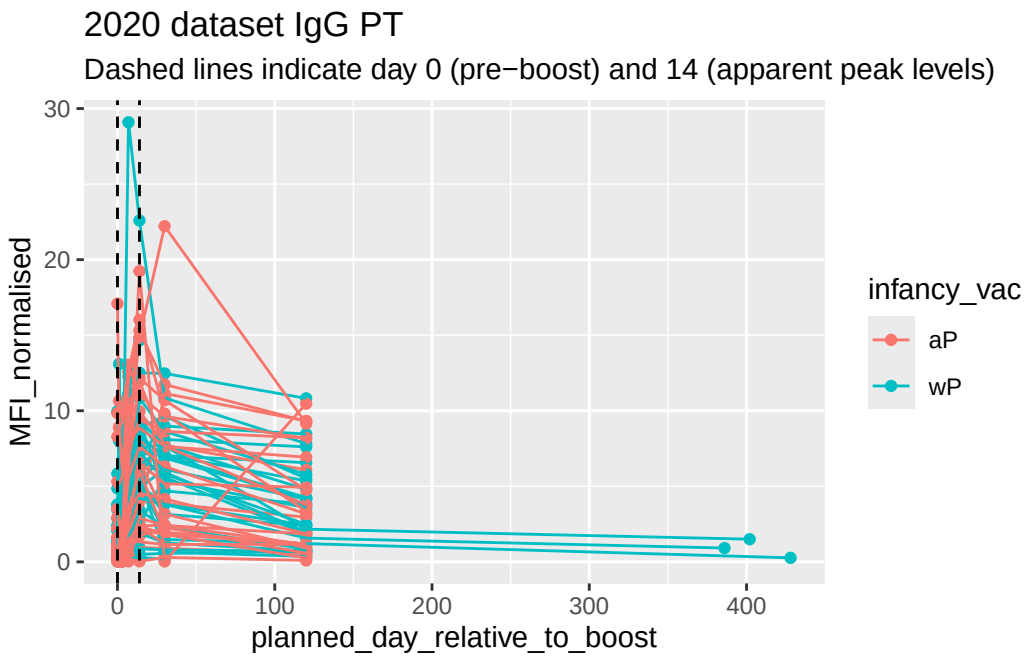


Let's finish here today, we are beginning to see some interesting difference between aP and wP individuals, There is likely lots of other interesting things to find in this dataset...

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

```
abdata.20 <- ab %>% 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)")
```



No, the trend on the 2021 dataset does not look similar to the 2020 dataset. The trend on the 2021 dataset looks completely different than the 2021.

##5. Obtaining CMI-PB RNASeq data

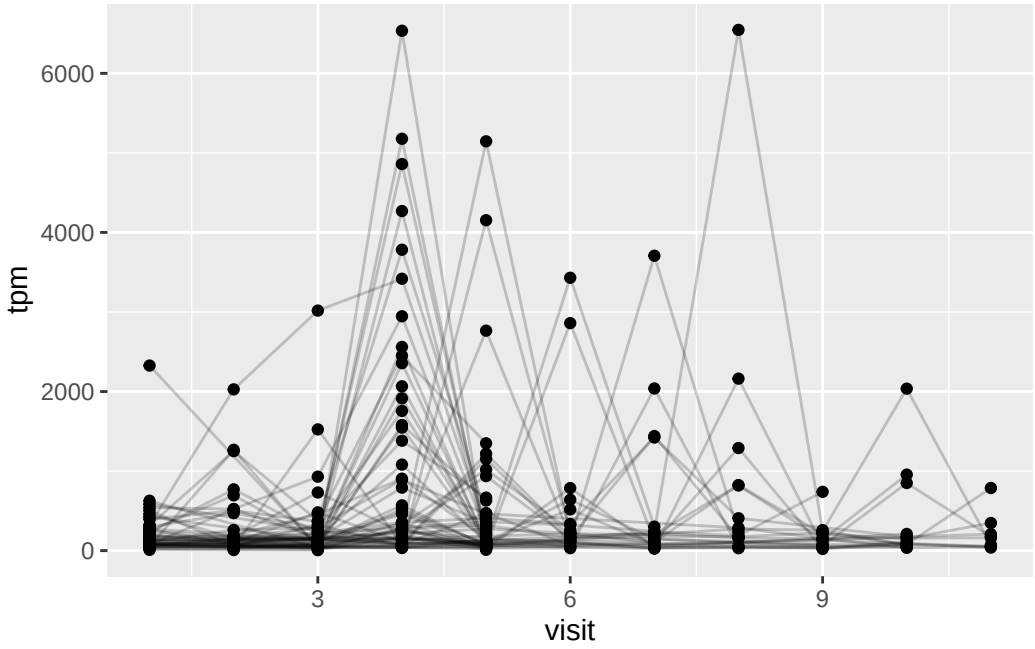
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

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)?

When the gene is expressed at its maximum level, there would always be a steep decline of tpm after each visit. This trend of inclining and declining pattern occurs throughout the plot.

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

This pattern matches the antibody titer data because there is a similar trend of PT levels rising over time, reaching to a peak, then declining. This trend is quite similar to the pattern that is shown in the the gene expression for IGHG1 plot.