# Class 18: Investigating Pertussis Resurgence

Snehita Vallumchetla (PID: A16853399)

## Table of contents

Pertussis (a.k.a) Whooping Cough is a deadly lung infection caused by the bacteria B. Pertussis.

The CDC tracks Pertussis cases around the US.

http://tinyurl.com/pertussiscdc

We can "scrape" this data using R **datapasta** package.

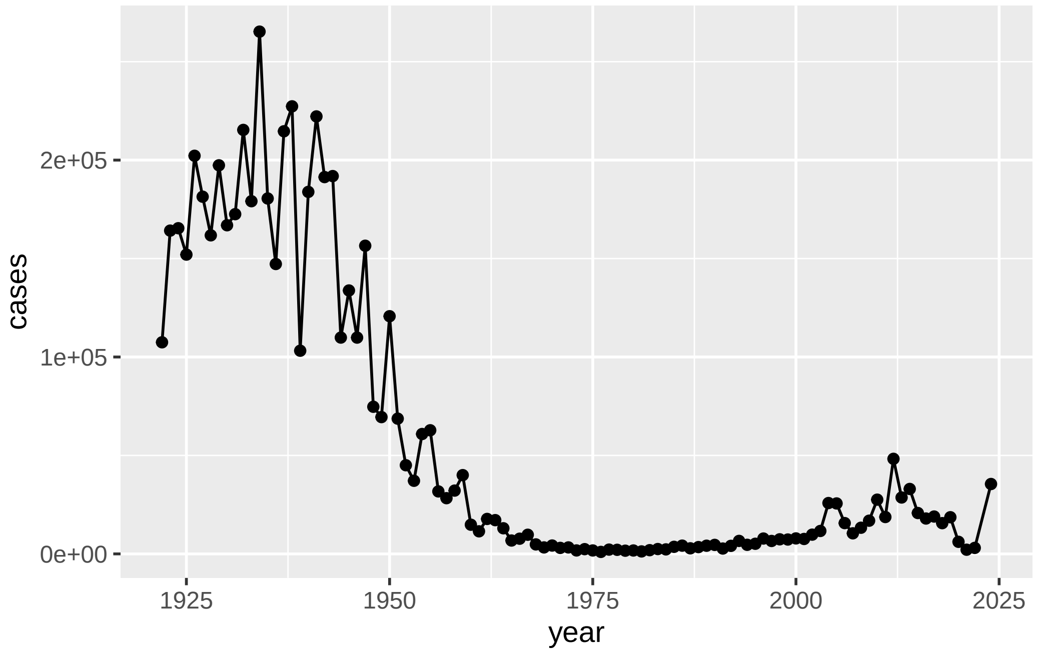
### 1. Investigating pertussis cases by year

> Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
cdc <- data.frame(
  year = c(
    1922L,
    1923L,
    1924L,
    1925L,
    1926L,
    1927L,
    1928L,
    1929L,
    1930L,
    1931L,
    1932L,
    1933L,
    1934L,
    1935L,
    1936L,
    1937L,
    1938L,
    1939L,
    1940L,
    1941L,
    1942L,
    1943L,
    1944L,
    1945L,
    1946L,
    1947L,
    1948L,
    1949L,
    1950L,
    1951L,
    1952L,
    1953L,
    1954L,
    1955L,
    1956L,
    1957L,
    1958L,
    1959L,
    1960L,
    1961L,
    1962L,
    1963L,
    1964L,
    1965L,
    1966L,
    1967L,
    1968L,
    1969L,
    1970L,
    1971L,
    1972L,
    1973L,
    1974L,
    1975L,
    1976L,
    1977L,
    1978L,
    1979L,
    1980L,
    1981L,
    1982L,
    1983L,
    1984L,
    1985L,
    1986L,
    1987L,
    1988L,
    1989L,
    1990L,
    1991L,
    1992L,
    1993L,
    1994L,
    1995L,
    1996L,
    1997L,
    1998L,
    1999L,
    2000L,
    2001L,
    2002L,
    2003L,
    2004L,
    2005L,
    2006L,
    2007L,
    2008L,
    2009L,
    2010L,
    2011L,
    2012L,
    2013L,
    2014L,
    2015L,
    2016L,
    2017L,
    2018L,
    2019L,
    2020L,
    2021L,
    2022L,
  2024L),
  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,
  35493)
)
```

```
library(ggplot2)
```

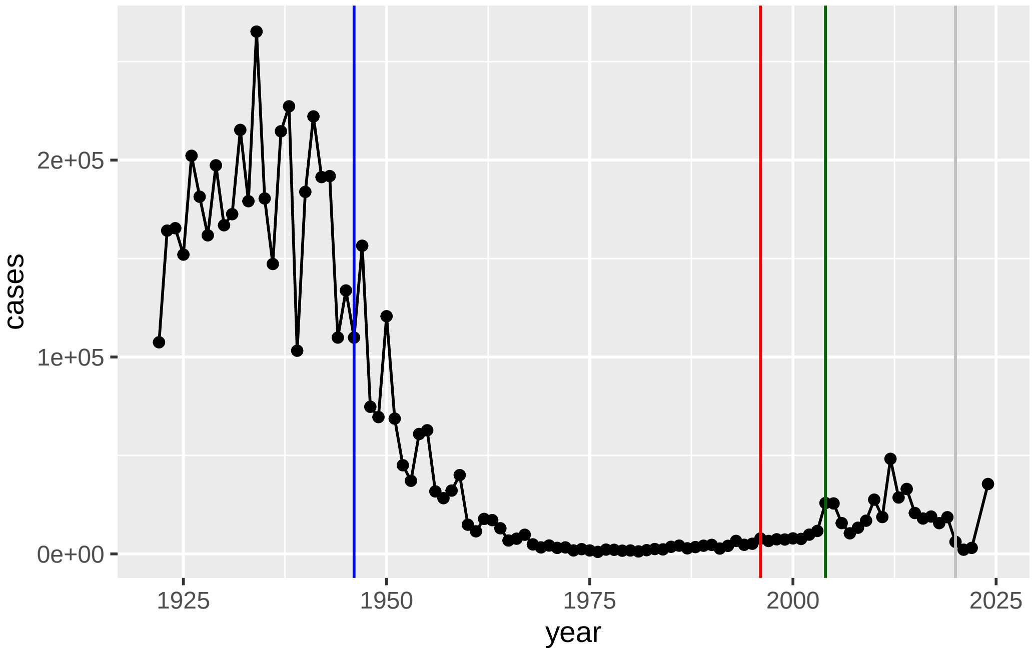
```
ggplot(data = cdc) +
  aes(x = year, y = cases) +
  geom_line() +
  geom_point()
```



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

Q2. Using the ggplot geom_vline() function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice?

```r
ggplot(data = cdc) +
  aes(x = year, y = cases) +
  geom_line() +
  geom_point() +
  geom_vline(xintercept = 1946, col = 'blue') +
  geom_vline(xintercept = 1996, col = 'red') +
  geom_vline(xintercept = 2020, col = 'grey') +
  geom_vline(xintercept = 2004, col = 'darkgreen')
```



There were high numbers before the first wP (whole-cell) vaccine roll out in 1946 then a rapid decline in case numbers until 2004 when we have our first large-scale outbreaks of pertussis again. There is also a notable COVID related dip and recent rapid rise.

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

There are some hypotheses for what happened after the aP vaccine which could be due to more sensitive PCR testing, and also due to vaccine hesitancy. Sources also say that there has been an evolution in bacteria, while the immunity of adolescents ents has been declining.

### 3. Computational Models of Immunity Pertussis Boost (CMI-PB)

The CMI-PB project aims to address this key question: what is different between aP and wP individuals.

We can get all the data from this ongoing project via JSON API calls. Fir this we will use the **jsonlite** package. We can install with `nstall.packages("jsonlite")`

```r
library(jsonlite)

subject <- read_json("https://www.cmi-pb.org/api/v5_1/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
```

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

There are 87 aP and 85 wP infancy vaccinated subjects in the dataset.

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

```

aP wP 
87 85 
```

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

There are 112 biological females and 60 biological males in this dataset

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

```

Female   Male 
   112     60 
```

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

```r
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
```

Obtain more data from CMI-PB:

```r
specimen <- read_json("https://www.cmi-pb.org/api/v5_1/specimen", simplifyVector = TRUE)

ab_data <- read_json("https://www.cmi-pb.org/api/v5_1/plasma_ab_titer", simplifyVector = TRU
```

```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(ab_data)
```

```
  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. 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:

I now have 3 tables of data from CMI-PB: `subject`, `specimen`, and `ab_data`. I need to join these tables so I will have all the info I need to work with.

For this we will use the `inner_join()` function from the **dplyr** package.

```r
library(dplyr)
```

```


Attaching package: 'dplyr'


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

    filter, lag


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

    intersect, setdiff, setequal, union
```

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

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

```r
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 specimen_id
1    1986-01-01    2016-09-12 2020_dataset           1
2    1986-01-01    2016-09-12 2020_dataset           2
3    1986-01-01    2016-09-12 2020_dataset           3
4    1986-01-01    2016-09-12 2020_dataset           4
5    1986-01-01    2016-09-12 2020_dataset           5
6    1986-01-01    2016-09-12 2020_dataset           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
```

```
dim(subject)
```

```
[1] 172   8
```

```
dim(specimen)
```

```
[1] 1503    6
```

```
dim(meta)
```

```
[1] 1503   13
```

> 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.

Now we can join our `ab_data` table to `meta` so we can

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

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

```r
head(abdata)
```

```
  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 specimen_id
1    1986-01-01    2016-09-12 2020_dataset           1
2    1986-01-01    2016-09-12 2020_dataset           1
3    1986-01-01    2016-09-12 2020_dataset           1
4    1986-01-01    2016-09-12 2020_dataset           1
5    1986-01-01    2016-09-12 2020_dataset           1
6    1986-01-01    2016-09-12 2020_dataset           1
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                           -3                             0         Blood
2                           -3                             0         Blood
3                           -3                             0         Blood
4                           -3                             0         Blood
5                           -3                             0         Blood
6                           -3                             0         Blood
  visit isotype is_antigen_specific antigen        MFI MFI_normalised  unit
1     1     IgE               FALSE   Total 1110.21154       2.493425 UG/ML
2     1     IgE               FALSE   Total 2708.91616       2.493425 IU/ML
3     1     IgG                TRUE      PT   68.56614       3.736992 IU/ML
4     1     IgG                TRUE     PRN  332.12718       2.602350 IU/ML
5     1     IgG                TRUE     FHA 1887.12263      34.050956 IU/ML
6     1     IgE                TRUE     ACT    0.10000       1.000000 IU/ML
  lower_limit_of_detection
1                 2.096133
2                29.170000
```

```
3                  0.530000
4                  6.205949
5                  4.679535
6                  2.816431
```

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

```
length(abdata$isotype)
```

```
[1] 61956
```

Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most "recent" dataset?

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

```

2020_dataset 2021_dataset 2022_dataset 2023_dataset 
       31520         8085         7301        15050 
```

There has been an increase in the number of rows for the most recent data set compared to 2021 and 2022, but still not as high as 2023.

```
table(abdata$antigen)
```

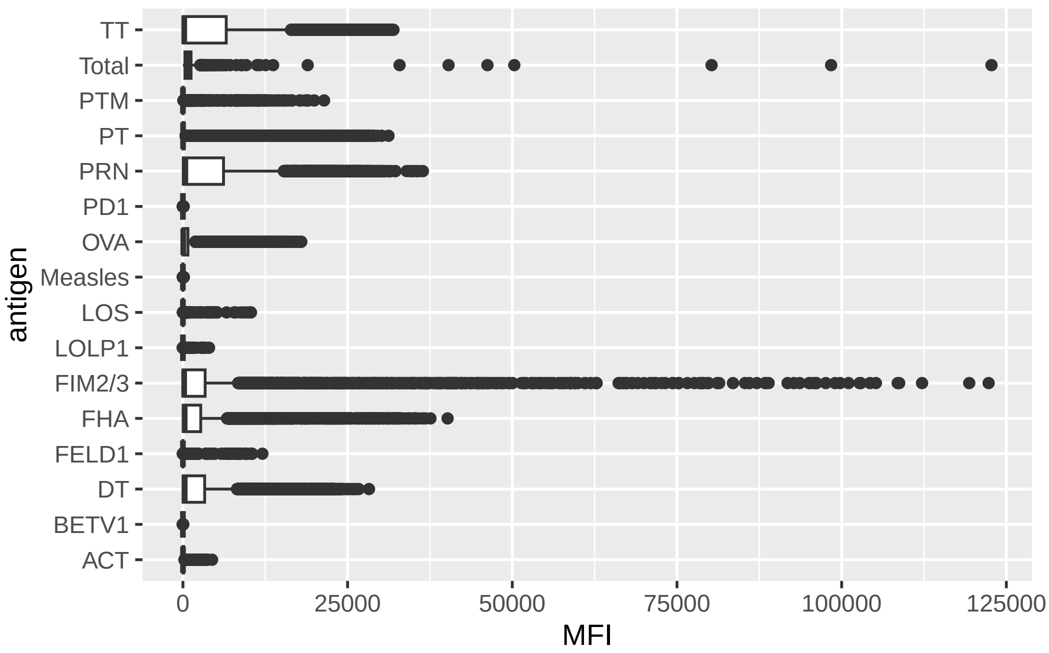
```

    ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA 
   1970    1970    6318    1970    6712    6318    1970    1970    1970    6318 
    PD1     PRN      PT     PTM   Total      TT 
   1970    6712    6712    1970     788    6318 
```

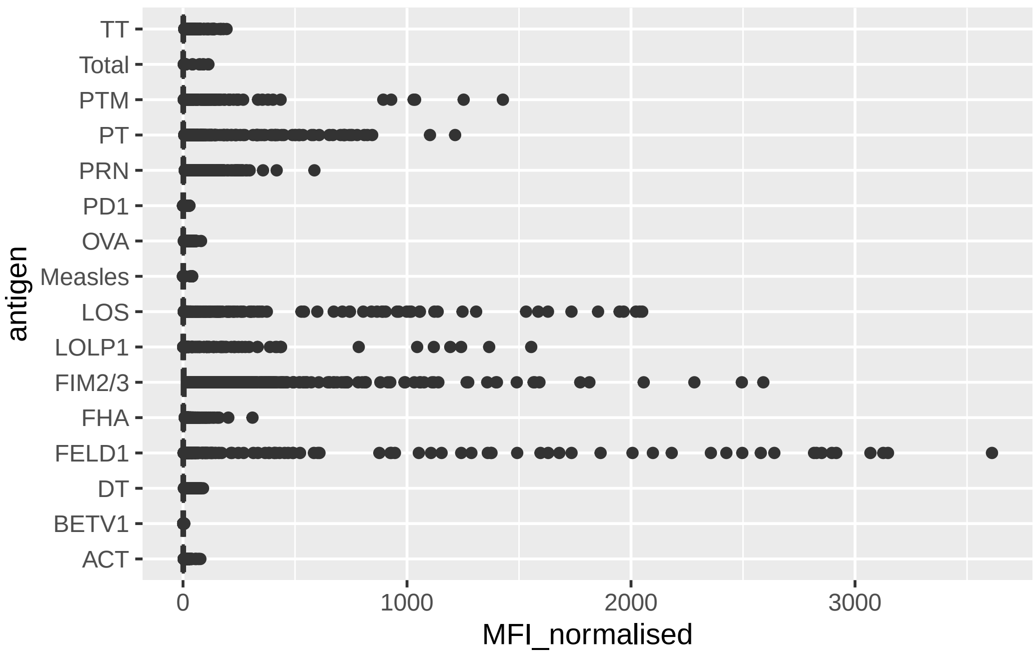
I want a plot of antigen levels across the whole dataset.

```
ggplot(abdata) +
  aes(MFI, antigen) +
  geom_boxplot()
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```



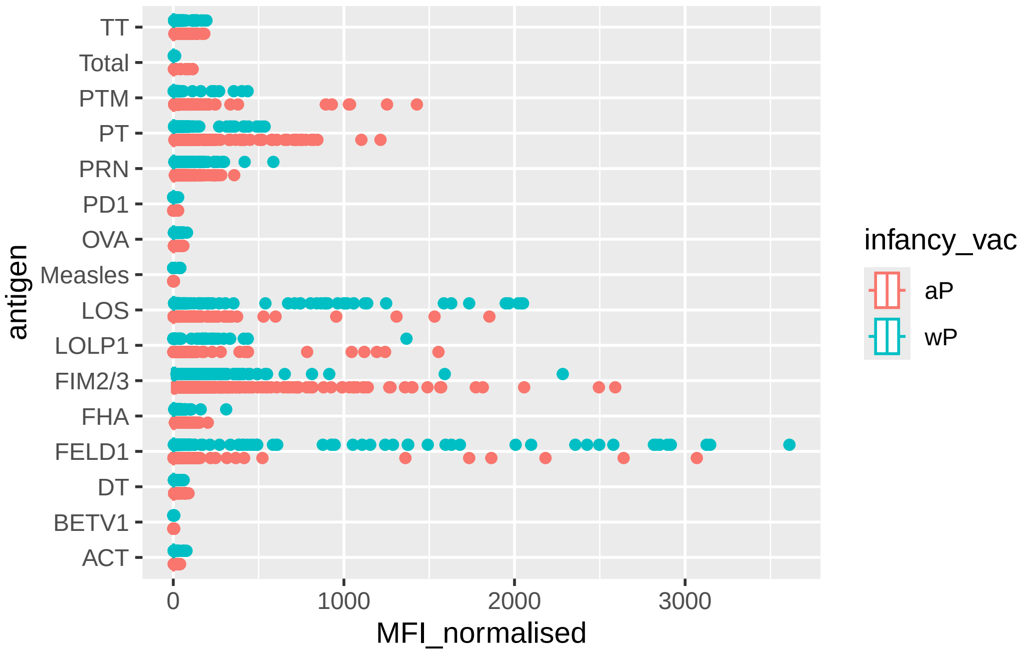
```
ggplot(abdata) +
  aes(MFI_normalised, antigen) +
  geom_boxplot()
```



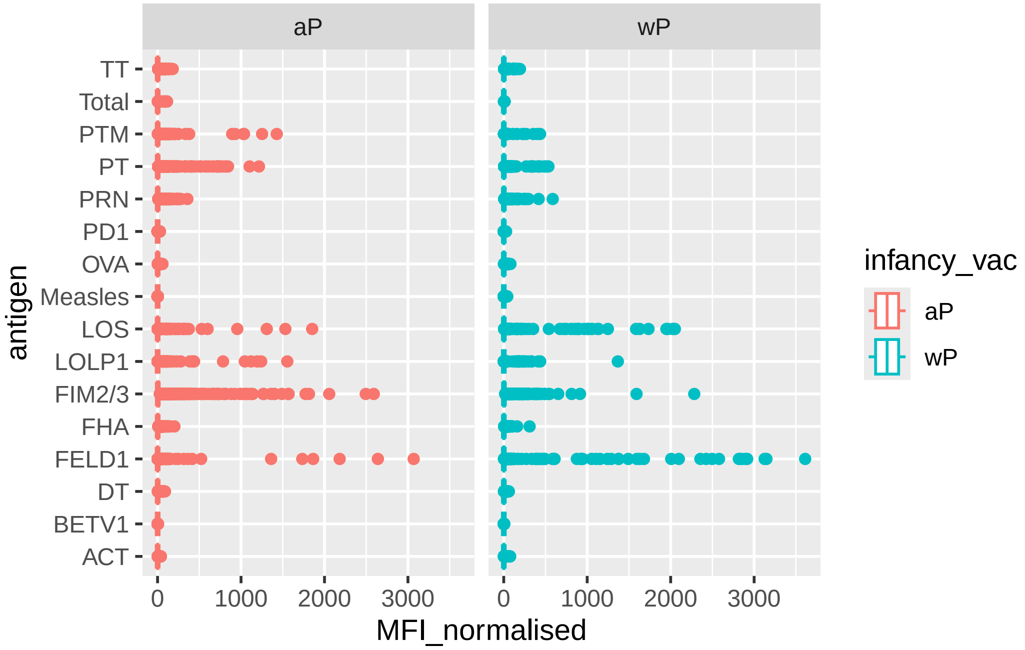
There are antigens like FIM2/3, PT, FELD1 have quite a large range of values. others like Measles dont show much activity.

> Q. Are there differences at this whole-dataset level between aP and wP?

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



```
ggplot(abdata) +
  aes(MFI_normalised, antigen, col = infancy_vac) +
  geom_boxplot() +
  facet_wrap(~infancy_vac)
```



## 4. Examine IgG Ab titer levels:

For this I need to select out just the isotype IgG

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

```
  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 specimen_id
1    1986-01-01    2016-09-12 2020_dataset           1
2    1986-01-01    2016-09-12 2020_dataset           1
3    1986-01-01    2016-09-12 2020_dataset           1
4    1986-01-01    2016-09-12 2020_dataset           2
5    1986-01-01    2016-09-12 2020_dataset           2
6    1986-01-01    2016-09-12 2020_dataset           2
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
```

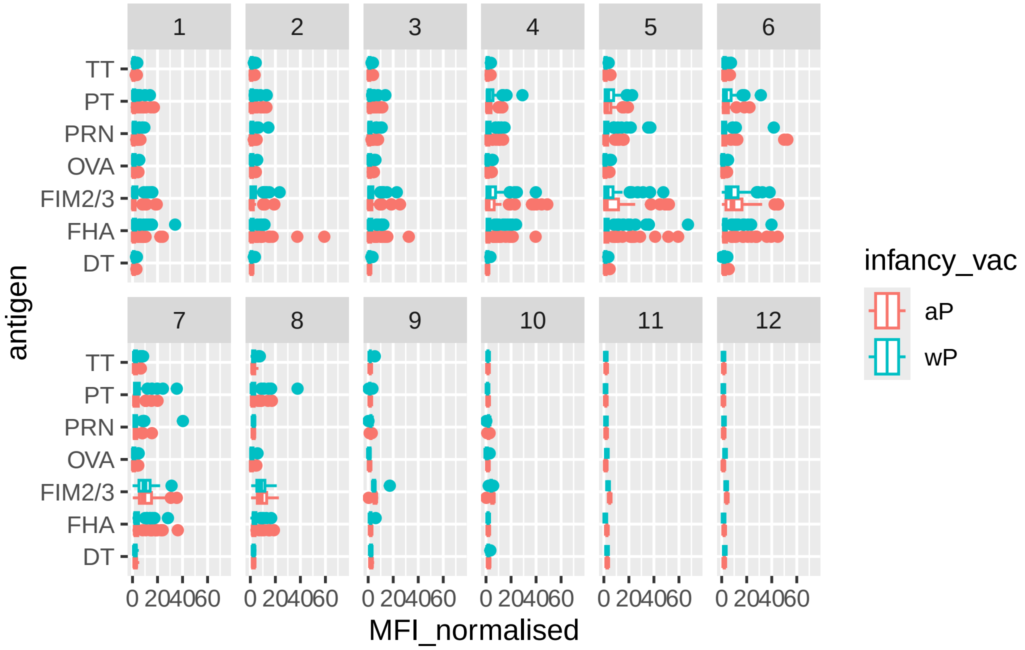
```
1                           -3                             0         Blood
2                           -3                             0         Blood
3                           -3                             0         Blood
4                            1                             1         Blood
5                            1                             1         Blood
6                            1                             1         Blood
  visit isotype is_antigen_specific antigen        MFI MFI_normalised  unit
1     1     IgG                TRUE      PT   68.56614       3.736992 IU/ML
2     1     IgG                TRUE     PRN  332.12718       2.602350 IU/ML
3     1     IgG                TRUE     FHA 1887.12263      34.050956 IU/ML
4     2     IgG                TRUE      PT   41.38442       2.255534 IU/ML
5     2     IgG                TRUE     PRN  174.89761       1.370393 IU/ML
6     2     IgG                TRUE     FHA  246.00957       4.438960 IU/ML
  lower_limit_of_detection
1                 0.530000
2                 6.205949
3                 4.679535
4                 0.530000
5                 6.205949
6                 4.679535
```

> Q13. Complete the following code to make a summary boxplot of Ab titer levels (MFI) for all antigens:

an overview boxplot

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

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

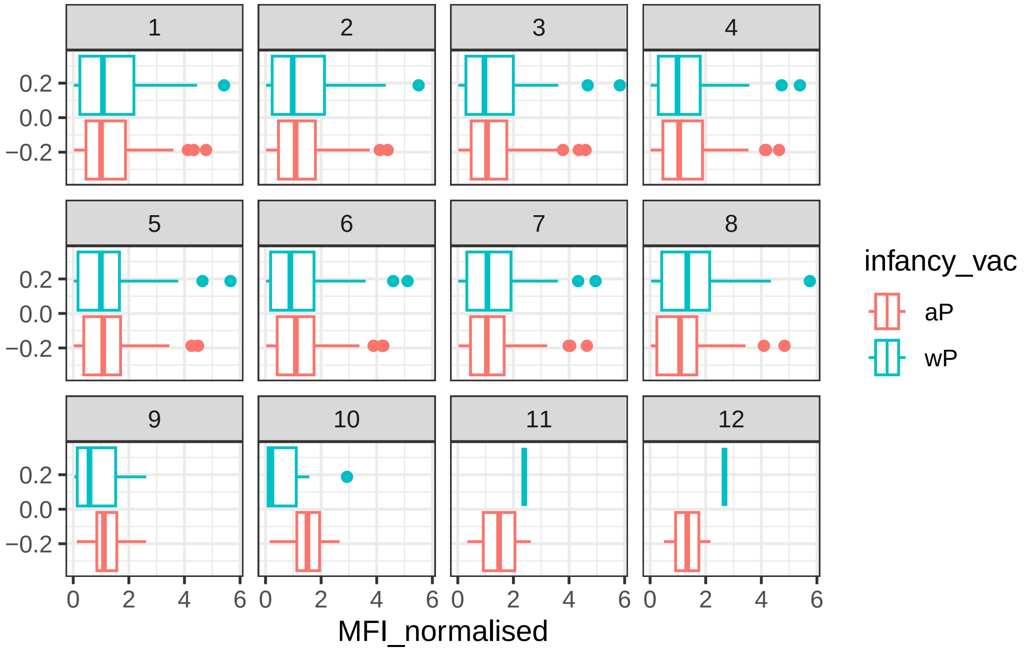


> Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

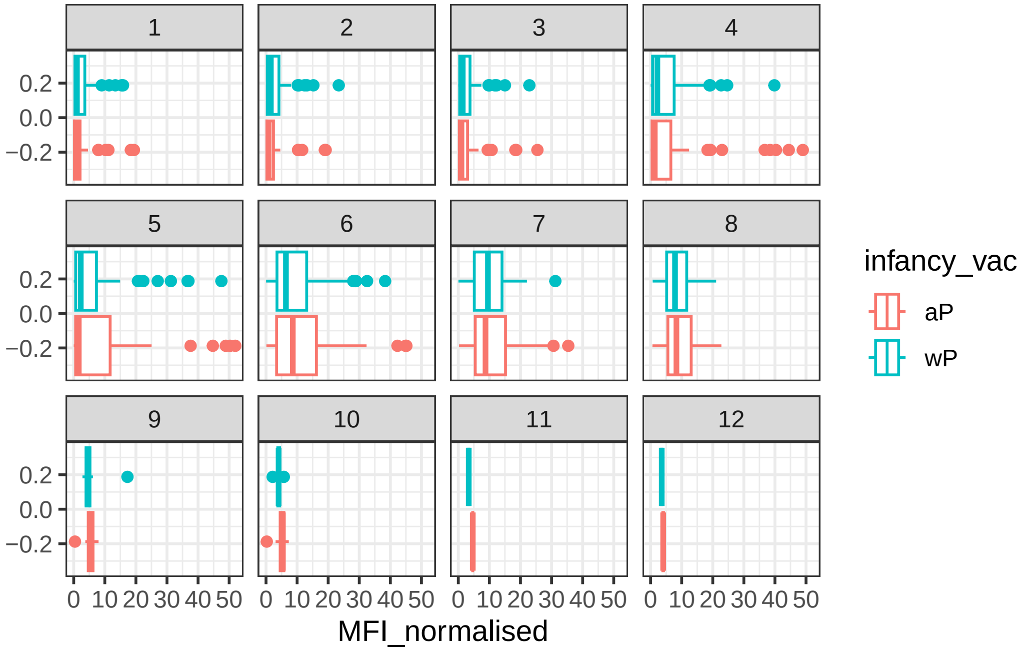
Some that show a difference over time are FHA and FIM2/3.

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

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



```
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?

Looking at the time courses, there seems to be a sharper decline of the FIM2/3 igg in comparison to the OVA, while there has been an overall decline in both.

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

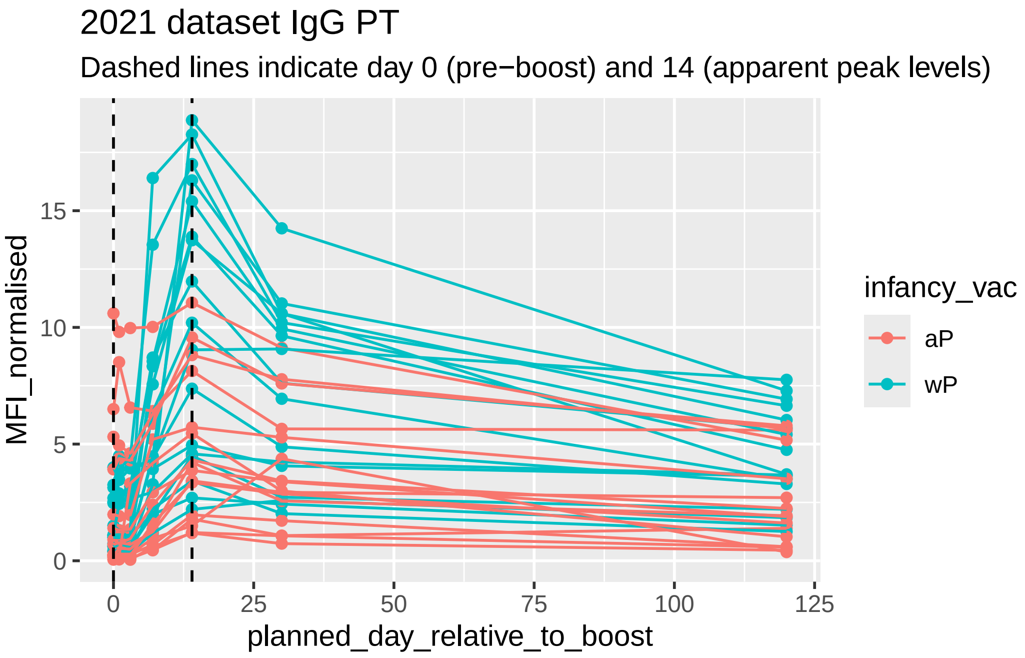
There seems to be a difference in response of aP and wP for the OVA antigen, whereas there is less of a difference in response to the aP and wP for the FIM2/3 antigen.

Digging in further to look at the time course of Igzg isotype PT antigen levels acrpss aP and wP individuals:

```
#filter to include 2021 daya only
abdata.21 <- abdata %>% filter(dataset == "2021_dataset")

# filter to look at IgG PT data only
abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT") %>%

# Plot and conlor by infancy_vac(wP vs aP)
  ggplot() +
    aes(x=planned_day_relative_to_boost,
        y=MFI_normalised,
        col=infancy_vac,
        group=subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
  labs(title="2021 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```



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

This trend looks similar to the 2020 dataset.

## 5. Obtaining CMI-PB RNASeq data

The link above is for the key gene involved in expressing any IgG1 antibody, namely the IGHG1 gene. Let's read available RNA-Seq data for this gene into R and investigate the time course of it's gene expression values.

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

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

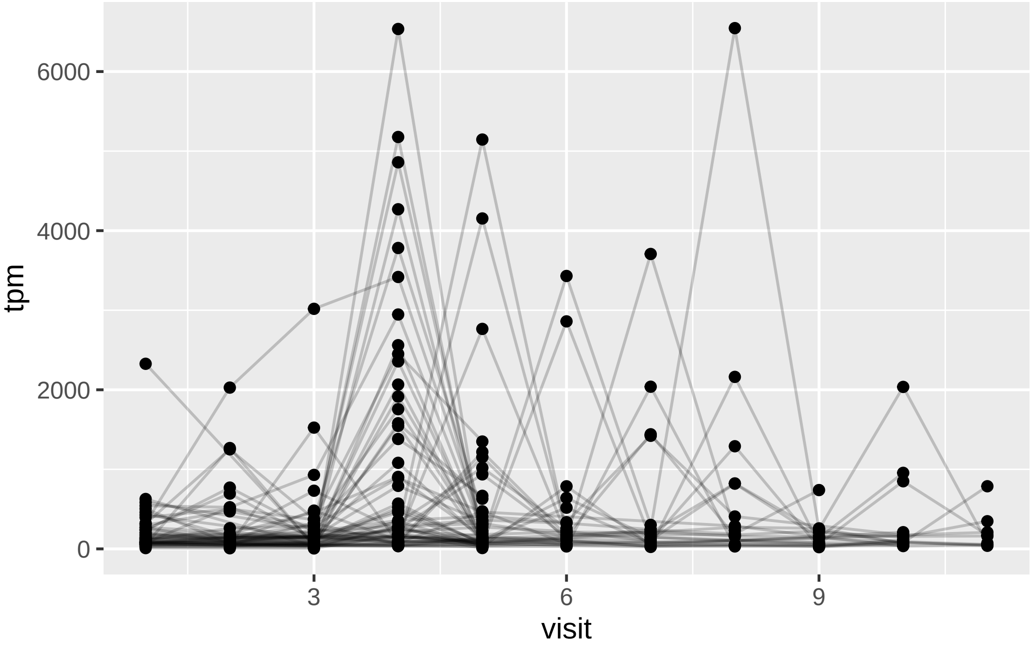
We will once again use the `innerjoin` function to join our metadata with the rna data to assist with further analysis:

```
#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)?

The gene is at its maximum level at visit 4 and at visit 8.

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

This trend does coincide with the antibody titer data, as there is an increase in spread (more outliers) during day 4 and 8.