Class 15: Investigating Pertussis Resurgence

Saba Heydari Seradj

Background

Pertussis (more commonly known as whooping cough) is a highly contagious respiratory disease caused by the bacterium Bordetella pertussis (Figure 1). People of all ages can be infected leading to violent coughing fits followed by a characteristic high-pitched "whoop" like intake of breath. Children have the highest risk for severe complications and death. Recent estimates from the WHO indicate that ~16 million cases and 200,000 infant deaths are due to pertussis annually (Black et al. 2010).

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.

Problem... data is in pdf format. We will use datapasta package to scrape data into R as a dataframe.

```
#install.packages('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,
```

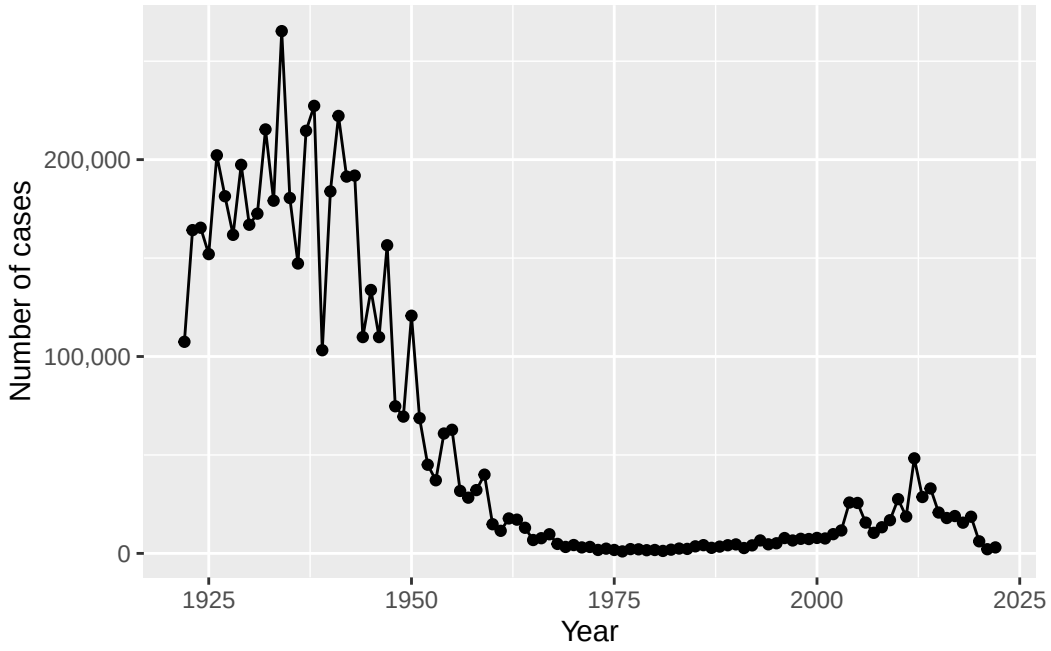
```
                                  1968L,1969L,1970L,1971L,1972L,1973L,
                                  1974L,1975L,1976L,1977L,1978L,1979L,1980L,
                                  1981L,1982L,1983L,1984L,1985L,1986L,
                                  1987L,1988L,1989L,1990L,1991L,1992L,1993L,
                                  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),
  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)
)
```

```
head(cdc)
```

```
  year  cases
1 1922 107473
2 1923 164191
3 1924 165418
4 1925 152003
5 1926 202210
6 1927 181411
```

```
library(ggplot2)
library(scales)
```

```
# Building the plot
baseplot <- ggplot(cdc) +
aes(x = year,
y = cases) +
geom_point() +
geom_line() +
labs(x = "Year",
y = "Number of cases") +
scale_y_continuous(labels = comma) # No longer scientific notation
baseplot
```

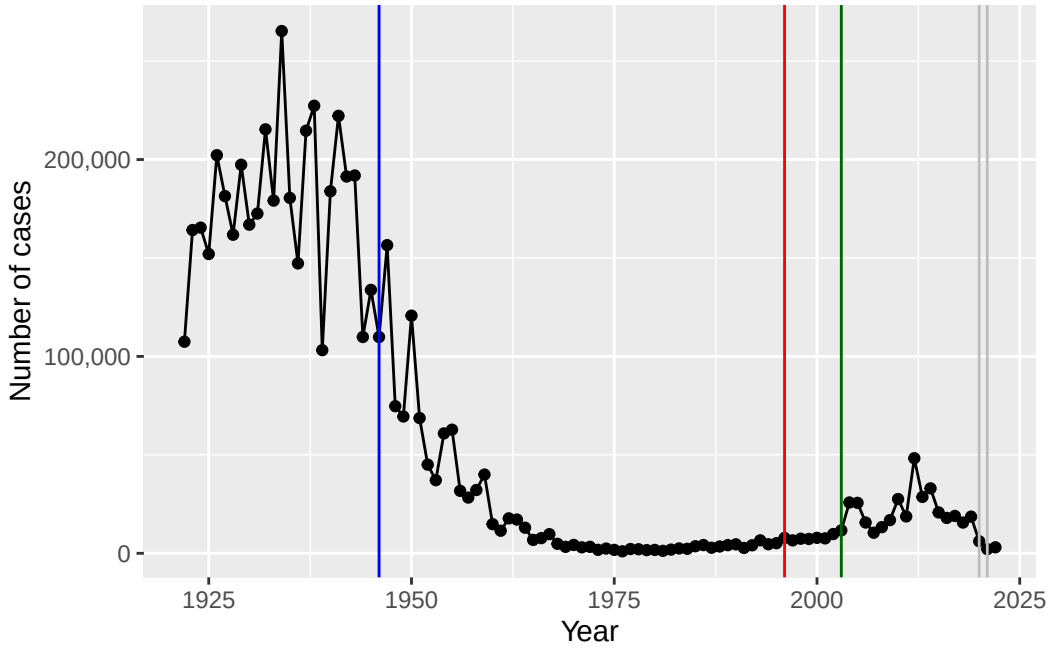


A tale of two vaccines (wP & aP)

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. What do you notice?

```
# Landmark plot
lm_plot <- baseplot +
geom_vline(xintercept = 1946, # wP vaccine with everything
```

```
col = 'blue') +
geom_vline(xintercept = 1996, # aP vaccine with "essential components"
col = 'red') +
geom_vline(xintercept = 2003, # Start of the big increase
col = 'darkgreen') +
geom_vline(xintercept = c(2020,2021), # Covid-19 lockdowns
col = 'grey')
lm_plot
```



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

aP was introduced in 1996 and for ~10 years there was a steady state. After that, around 2004, there was a big increase in cases which could be due to various reasons, such as bacterial evolution, short-lasting effectiveness of aP, and antivax movements.

Exploring CMI-PB data

Problem...data is in JSON format. We will use jsonlite package to process JSON data.

```
#install.packages('jsonlite')
```

```
library(jsonlite)

# Read subject table
subject <- read_json('https://www.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
```

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

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

```
aP wP
87 85
```

There are 85 wP and 87 aP vaccinated subjects.

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

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

```
Female   Male
   112     60
```

112 females, 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
```

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:

```
specimens <- read_json('https://www.cmi-pb.org/api/v5/specimen',
simplifyVector = TRUE)
head(specimens)
```

```
  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
```

```
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
```

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

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

```
head(meta)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                            1
3           3          1                            3
4           4          1                            7
5           5          1                           11
6           6          1                           32
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                             1         Blood     2          wP         Female
```

```
3                            3         Blood     3          wP         Female
4                            7         Blood     4          wP         Female
5                           14         Blood     5          wP         Female
6                           30         Blood     6          wP         Female
                ethnicity  race year_of_birth date_of_boost      dataset
1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
```

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.

```
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
```

```
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 actual_day_relative_to_boost
1 UG/ML                2.096133          1                           -3
2 IU/ML               29.170000          1                           -3
3 IU/ML                0.530000          1                           -3
4 IU/ML                6.205949          1                           -3
5 IU/ML                4.679535          1                           -3
6 IU/ML                2.816431          1                           -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 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
```

```
nrow(ab)
```

```
[1] 52576
```

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
```

Number of 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
```

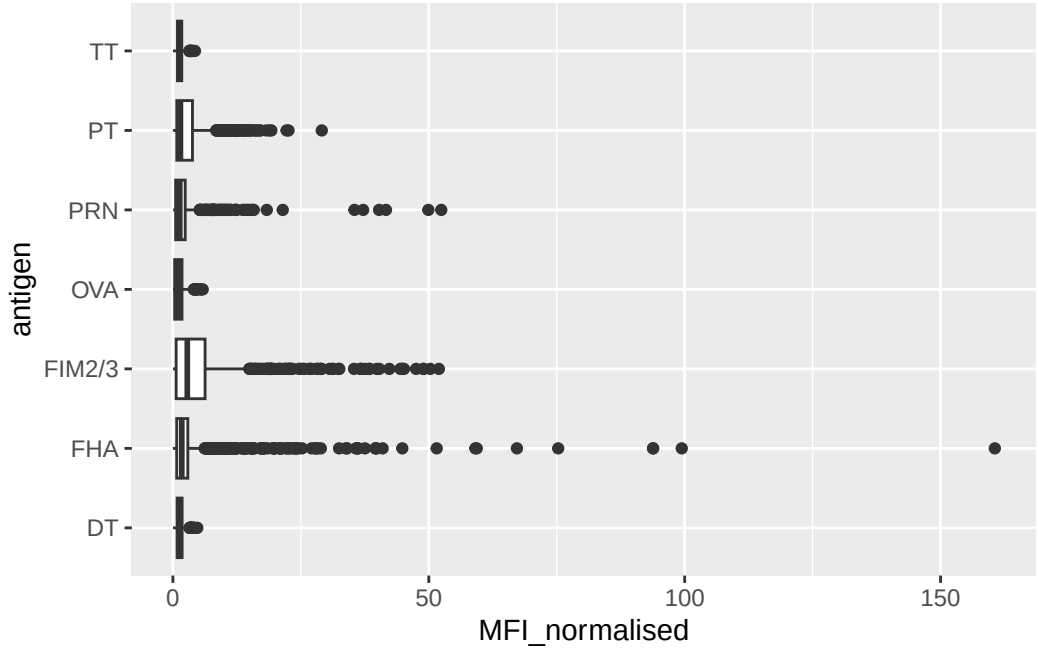
Focusing on IgG...

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

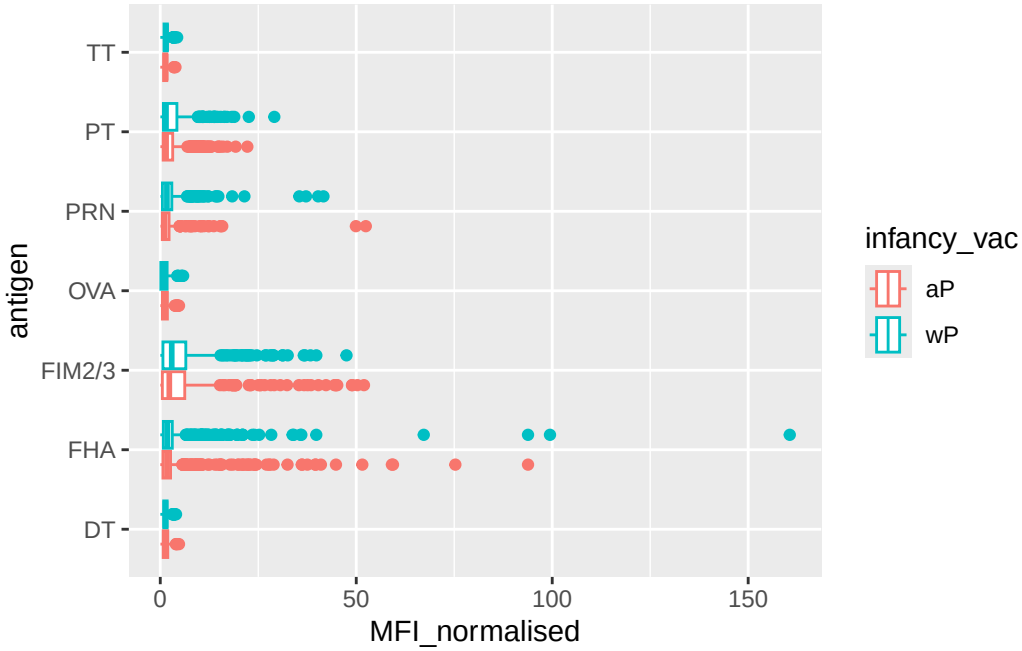
```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
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 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
```

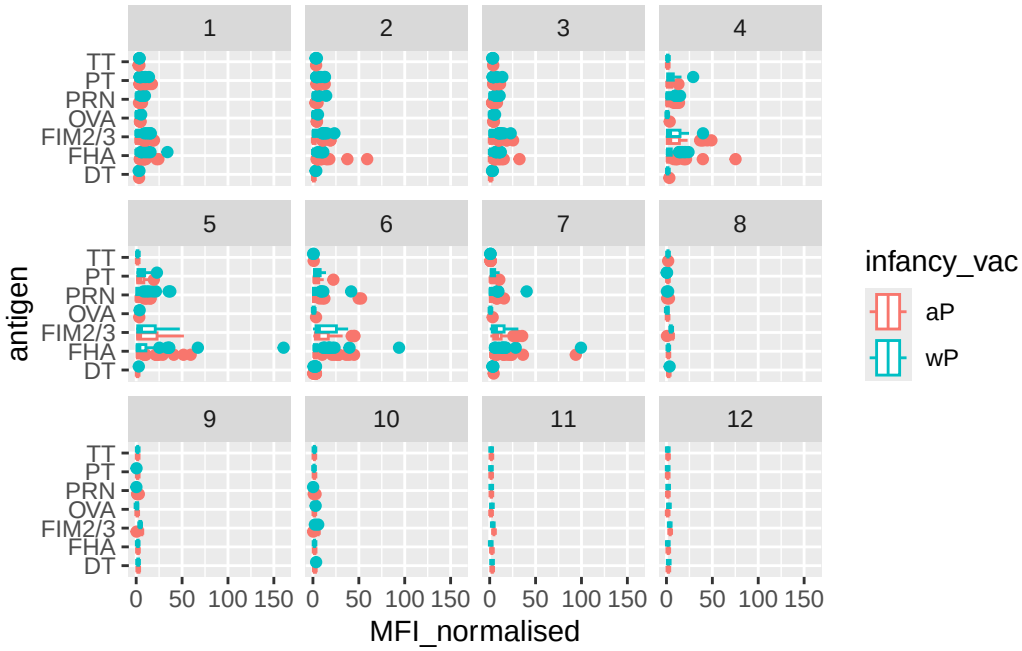
```
ggplot(igg) +
aes(MFI_normalised, antigen) +
geom_boxplot()
```



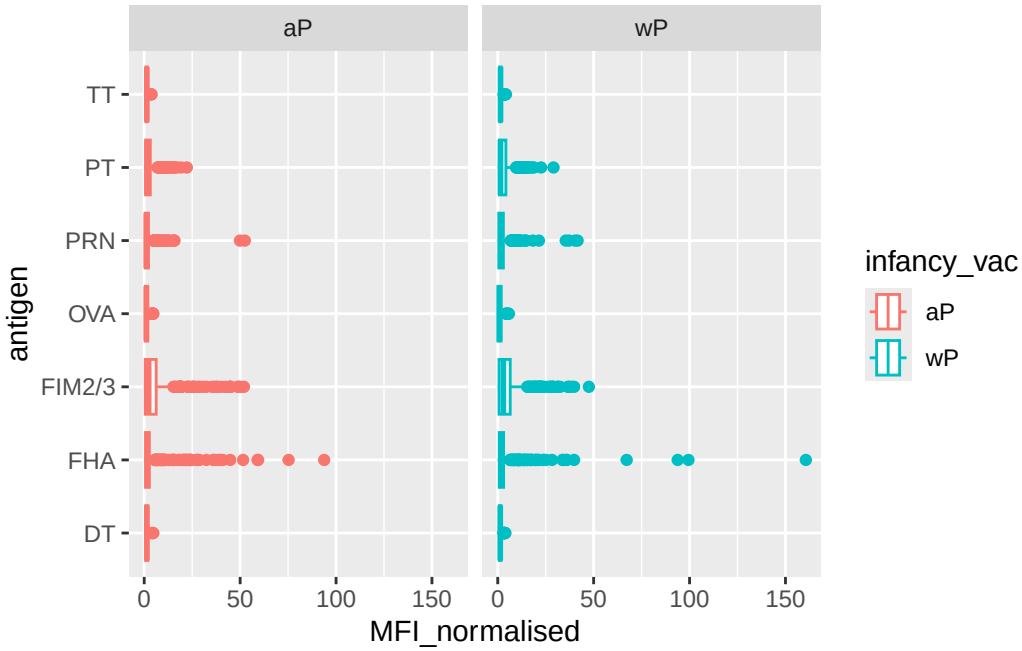

```
ggplot(igg) +
aes(MFI_normalised, antigen, col = infancy_vac) +
geom_boxplot()
```




```
ggplot(igg) +
aes(MFI_normalised, antigen, col = infancy_vac) +
geom_boxplot() +
facet_wrap(~visit) # Faceting by visit
```




```
ggplot(igg) +
aes(MFI_normalised, antigen, col = infancy_vac) +
geom_boxplot() +
facet_wrap(~infancy_vac) # Faceting by vaccine
```




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

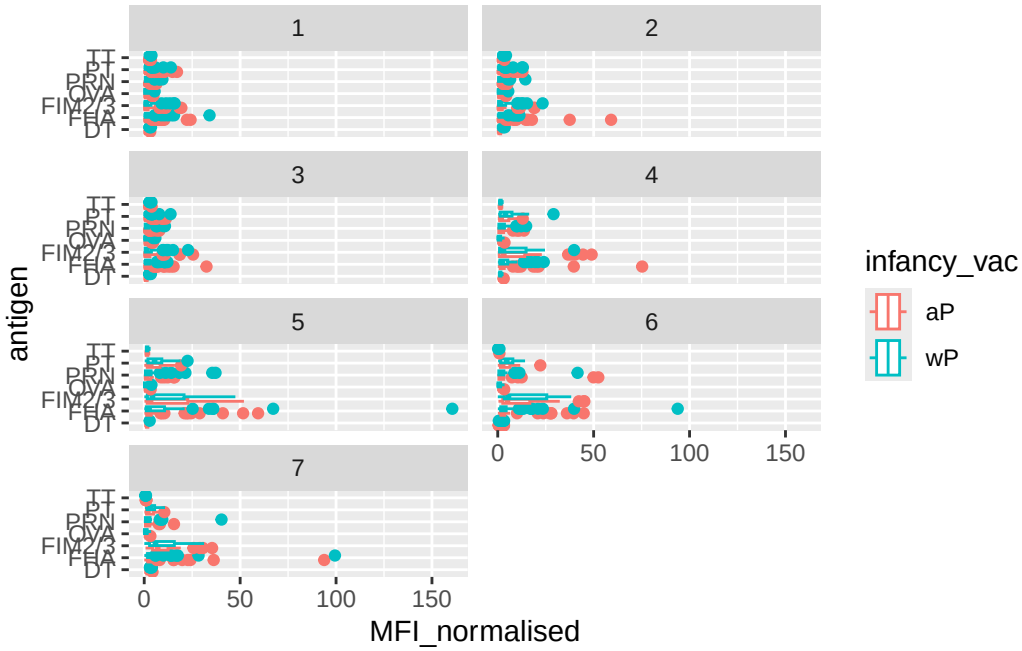
```
  1   2   3   4   5   6   7   8   9  10  11  12
902 902 930 559 559 540 525 150 147 133  21  21
```

We can see a trend. THere are a lot of visitations in the beginning but decreases towards later ones. We'll focus on visits 1-7.

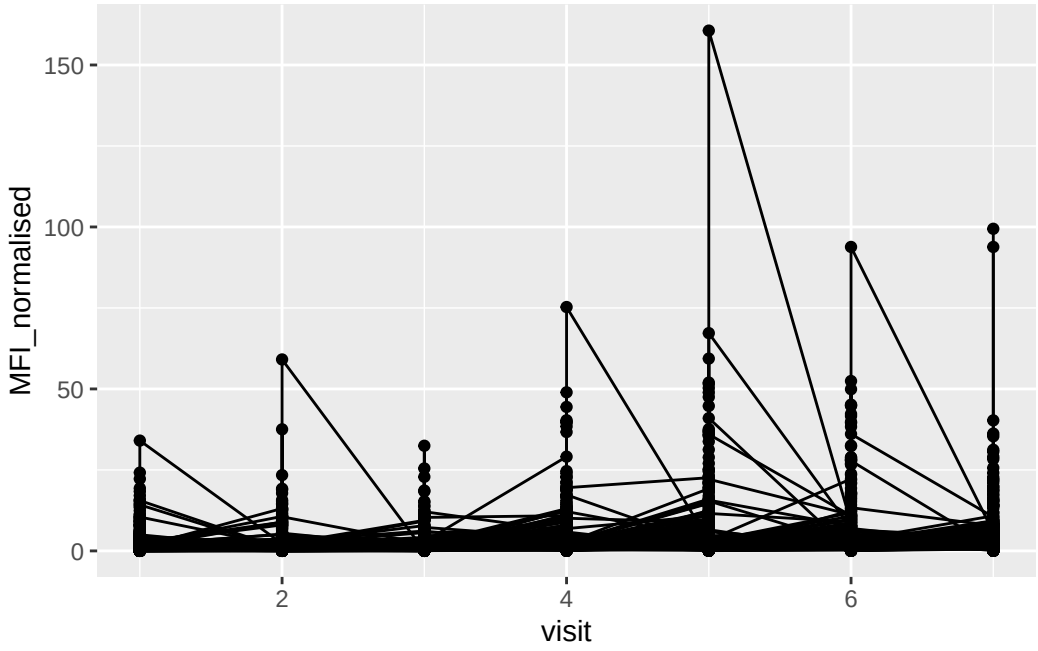
```
igg_7 <- filter(igg, visit %in% 1:7)
table(igg_7$visit)
```

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

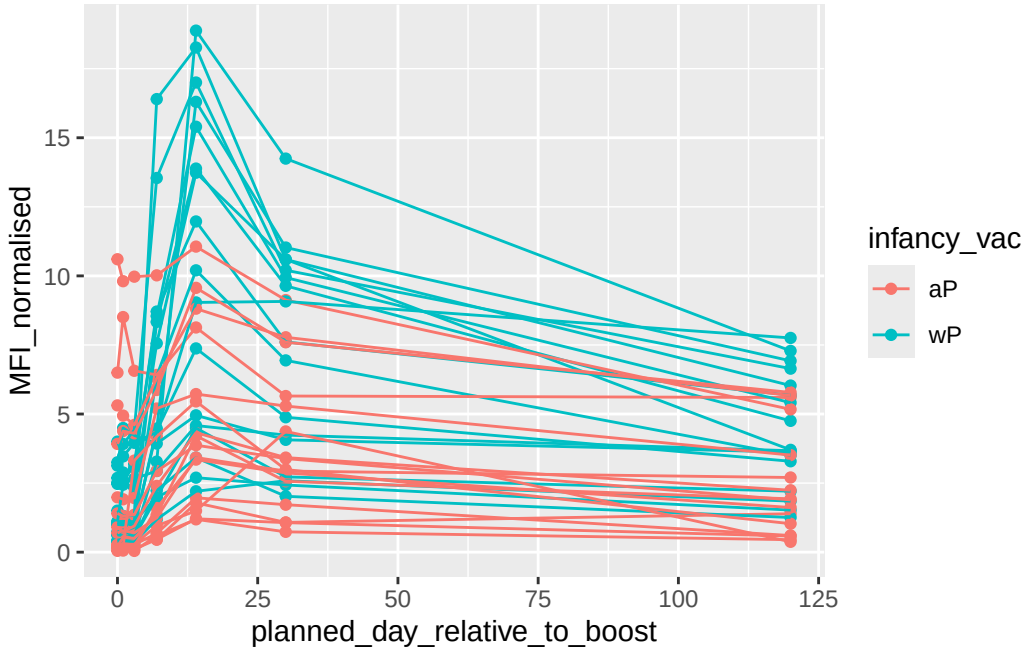
```
ggplot(igg_7) +
aes(MFI_normalised, antigen, col = infancy_vac) +
geom_boxplot() +
facet_wrap(~visit, ncol = 2) # Faceting by visit
```




```
ggplot(igg_7) +
aes(x = visit,
y = MFI_normalised,
group=subject_id) +
geom_point() +
geom_line()
```




```
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()
```


MFI_normalised
15
10
5
0
0
25
50
75
100
125
planned_day_relative_to_boost
infancy_vac
aP
wP