

# Class 17: Investigating Pertussis

Peter Shamasha (A15857589)

Pertussis, or whooping cough, is a highly contagious lung infection caused by a bacteria *B. pertussis*

The CDC tracks reported cases in the U.S. since the 1920s

```
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),
  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,
```

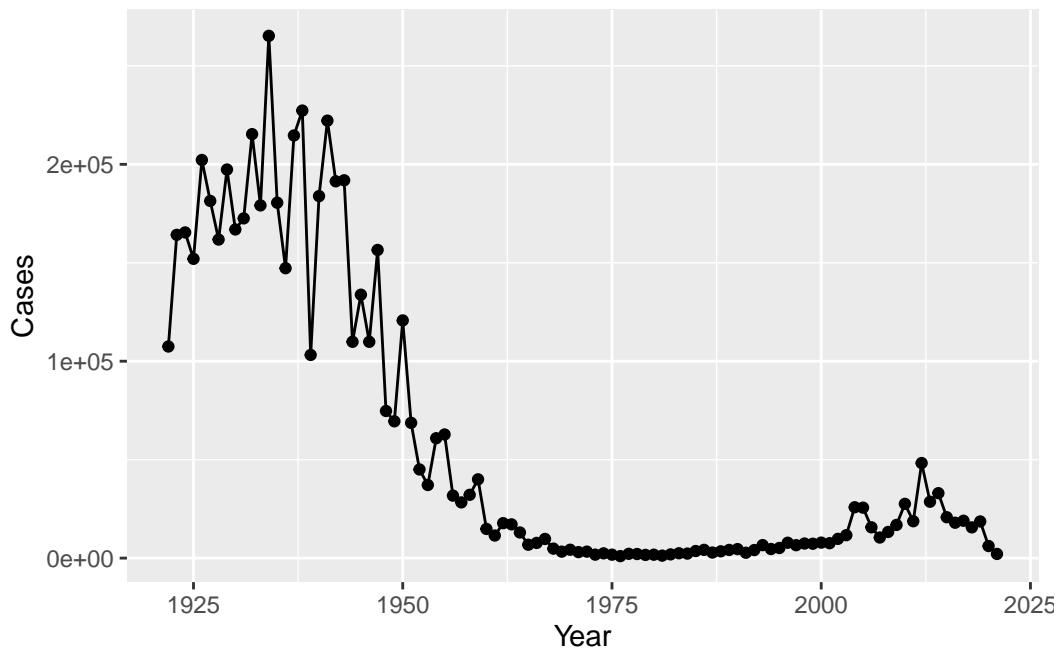
```
62786,31732,28295,32148,40005,
14809,11468,17749,17135,13005,6799,
7717,9718,4810,3285,4249,3036,
3287,1759,2402,1738,1010,2177,2063,
1623,1730,1248,1895,2463,2276,
3589,4195,2823,3450,4157,4570,
2719,4083,6586,4617,5137,7796,6564,
7405,7298,7867,7580,9771,11647,
25827,25616,15632,10454,13278,
16858,27550,18719,48277,28639,32971,
20762,17972,18975,15609,18617,
6124,2116)
```

```
)
```

We can now plot the number of reported pertussis cases per year in the U.S

```
library(ggplot2)

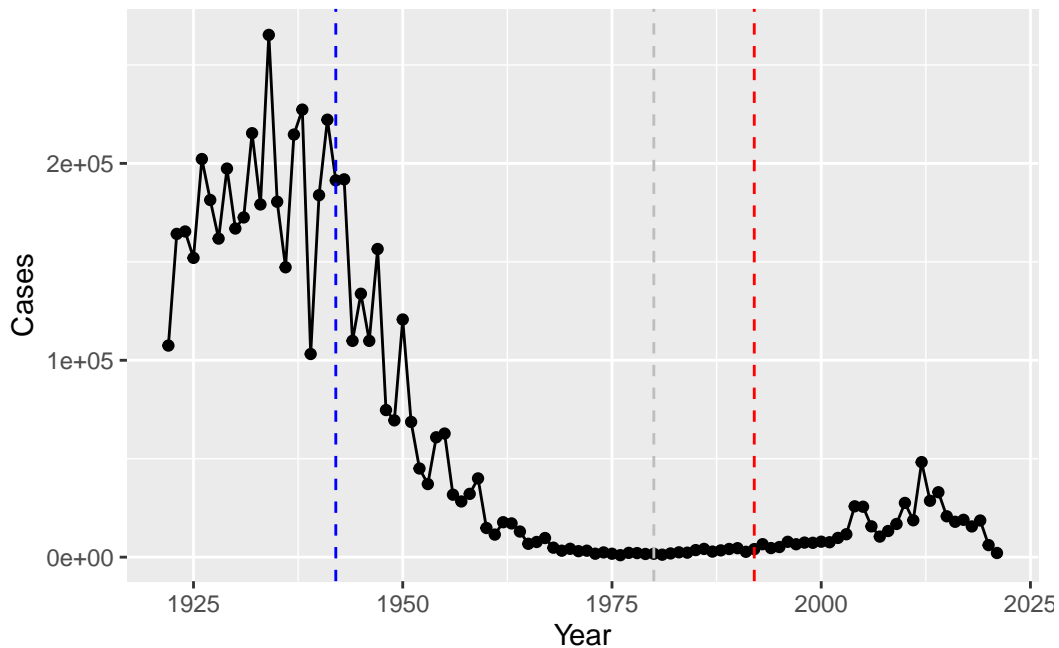
ggplot(cdc) +
  aes(Year, Cases) +
  geom_point() +
  geom_line()
```



The first big “whole-cell” pertussis vaccine program started in 1942

```
library(ggplot2)

ggplot(cdc) +
  aes(Year, Cases) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1942, col = "blue", linetype = 2) +
  geom_vline(xintercept = 1980, col = "gray", linetype = 2) +
  geom_vline(xintercept = 1992, col = "red", linetype = 2)
```



Something big is happening with pertussis cases and big outbreaks are once again a major public health concern!

One of the main hypothesis for the increasing case numbers is waning vaccine efficiency with the newer aP vaccine.

Enter the CMI-PB project, which is studying this problem on a large scale. Let’s see what data they have.

Their data is available in JSON format (“key:value” pair style). We will use the “jsonlite” package to read their data.

```
library(jsonlite)
```

```
subject <- read_json("https://www.cmi-pb.org/api/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
47 49
```

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

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

```
Female  Male
66      30
```

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                                     | 18     | 9    |
| Black or African American                 | 2      | 0    |
| More Than One Race                        | 8      | 2    |
| Native Hawaiian or Other Pacific Islander | 1      | 1    |
| Unknown or Not Reported                   | 10     | 4    |
| White                                     | 27     | 13   |

Now Let's read some more database tables from CMI-PB:

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

|   | specimen_id                   | subject_id    | actual_day_relative_to_boost |     |
|---|-------------------------------|---------------|------------------------------|-----|
| 1 | 1                             | 1             |                              | -3  |
| 2 | 2                             | 1             |                              | 736 |
| 3 | 3                             | 1             |                              | 1   |
| 4 | 4                             | 1             |                              | 3   |
| 5 | 5                             | 1             |                              | 7   |
| 6 | 6                             | 1             |                              | 11  |
|   | planned_day_relative_to_boost | specimen_type | visit                        |     |
| 1 | 0                             | Blood         | 1                            |     |
| 2 | 736                           | Blood         | 10                           |     |
| 3 | 1                             | Blood         | 2                            |     |
| 4 | 3                             | Blood         | 3                            |     |
| 5 | 7                             | Blood         | 4                            |     |
| 6 | 14                            | Blood         | 5                            |     |

I want to “join” (a.k.a. “merge”/link/etc.) the `subject` and `specimen` tables together. I will use the **dplyr** package for this.

```
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(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      | 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 | 736                          | 736                           | Blood         |
| 3 | 1                            | 1                             | Blood         |
| 4 | 3                            | 3                             | Blood         |
| 5 | 7                            | 7                             | Blood         |
| 6 | 11                           | 14                            | Blood         |

|   | visit |
|---|-------|
| 1 | 1     |
| 2 | 10    |
| 3 | 2     |
| 4 | 3     |
| 5 | 4     |
| 6 | 5     |

```
ab <- read_json("http://cmi-pb.org/api/ab_titer", simplifyVector = T)
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 |
|---|-------|--------------------------|
| 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                 |

Now I can join “meta” that we made above and contains all info about the subjects and specimens with this ab data.

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

Joining with `by = join\_by(specimen\_id)`

```
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

```

```
dim(abdata)
```

```
[1] 32675    20
```

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

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

```

IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 1413 6141 6141 6141 6141

```

Q12. What do you notice about the number of visit 8 specimens compared to other visits?

```
table(abdata$visit)
```



|      |      |      |      |      |      |      |    |
|------|------|------|------|------|------|------|----|
| 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8  |
| 5795 | 4640 | 4640 | 4640 | 4640 | 4320 | 3920 | 80 |

There are way less visit 8 specimens because the project is still ongoing and we have not gotten that data for all individuals yet.

## Examin IgG1 Ab titer levels

We will use the `filter()` function from `dplyr` to focus on just IgG1 isotype and visits 1 to 7 (i.e. exclude visit 8 as there are not many specimens there yet).

```
ig1 <- filter(abdata, isotype == "IgG1", visit!=8)
head(ig1)
```

|   | 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                            | IgG1                          | TRUE                | ACT                    | 274.355068 | 0.6928058      | IU/ML |
| 2 | 1                            | IgG1                          | TRUE                | LOS                    | 10.974026  | 2.1645083      | IU/ML |
| 3 | 1                            | IgG1                          | TRUE                | FELD1                  | 1.448796   | 0.8080941      | IU/ML |
| 4 | 1                            | IgG1                          | TRUE                | BETV1                  | 0.100000   | 1.0000000      | IU/ML |

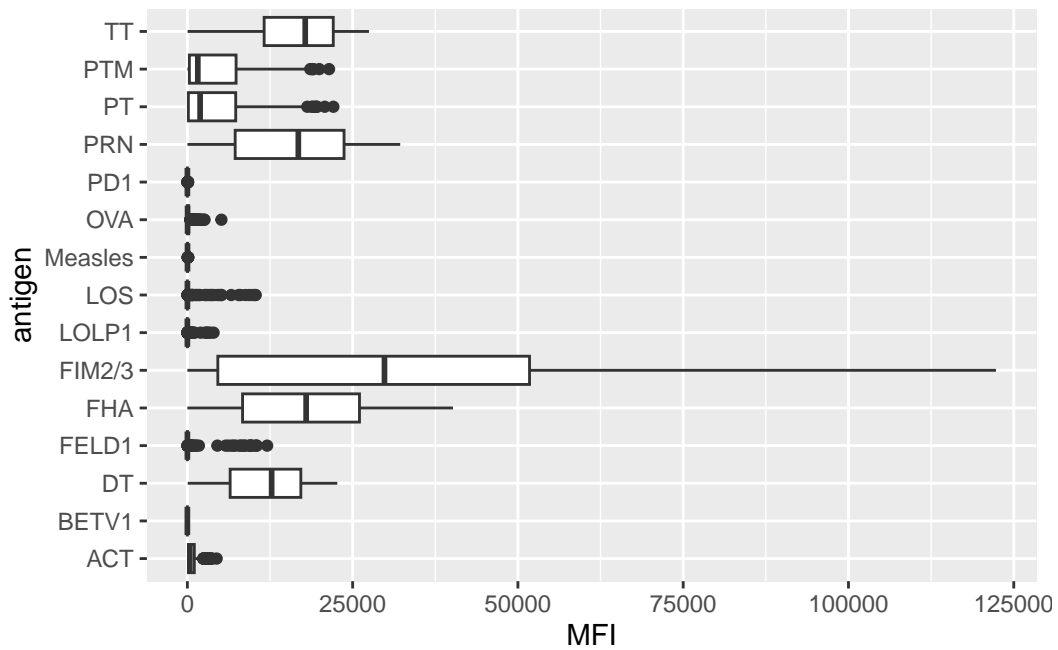
|   |   |      |      |         |           |                 |
|---|---|------|------|---------|-----------|-----------------|
| 5 | 1 | IgG1 | TRUE | LOLP1   | 0.100000  | 1.0000000 IU/ML |
| 6 | 1 | IgG1 | TRUE | Measles | 36.277417 | 1.6638332 IU/ML |

|   | lower_limit_of_detection |
|---|--------------------------|
| 1 | 3.848750                 |
| 2 | 4.357917                 |
| 3 | 2.699944                 |
| 4 | 1.734784                 |
| 5 | 2.550606                 |
| 6 | 4.438966                 |

Box plot of antigen levels over time.

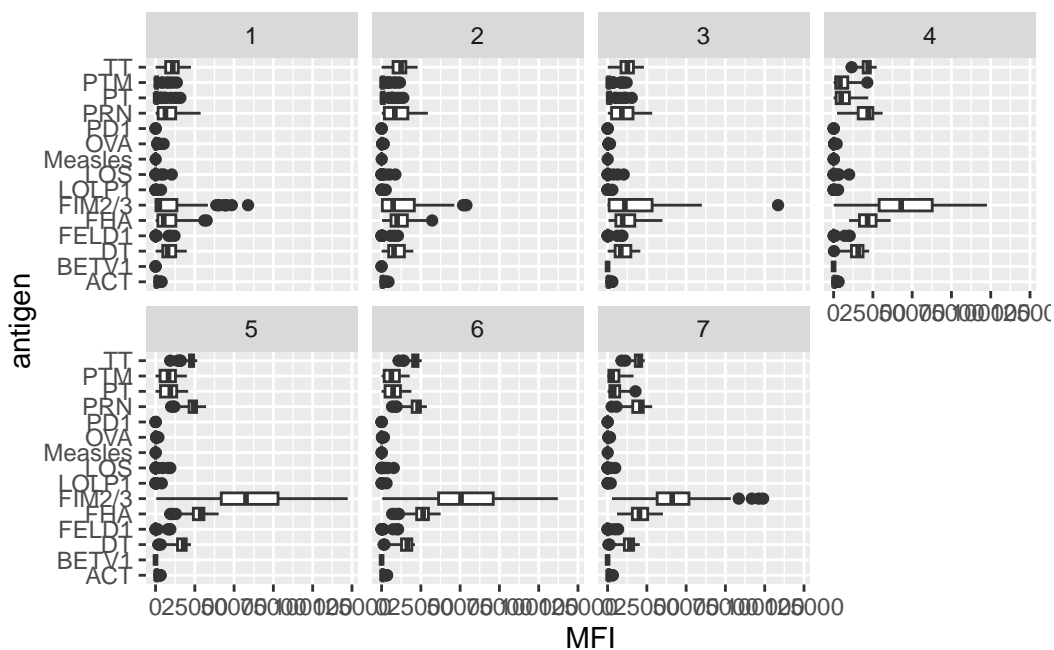
```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot()
```



And facet by visit

```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
```

```
facet_wrap(vars(visit), nrow=2)
```



Clearly FIM2/3 changes. This is “Fimbrial protein” that makes the bacteria plius and is involved in cell adhesion.

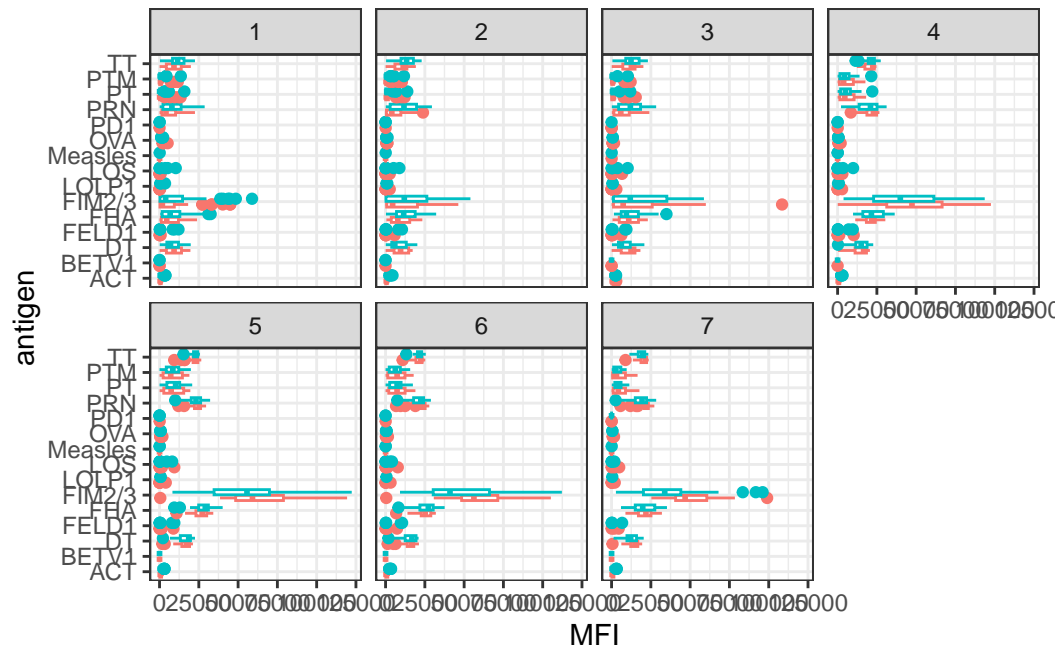
PT Pertussis toxin

FHA is Filamentous hemagglutinin

etc.

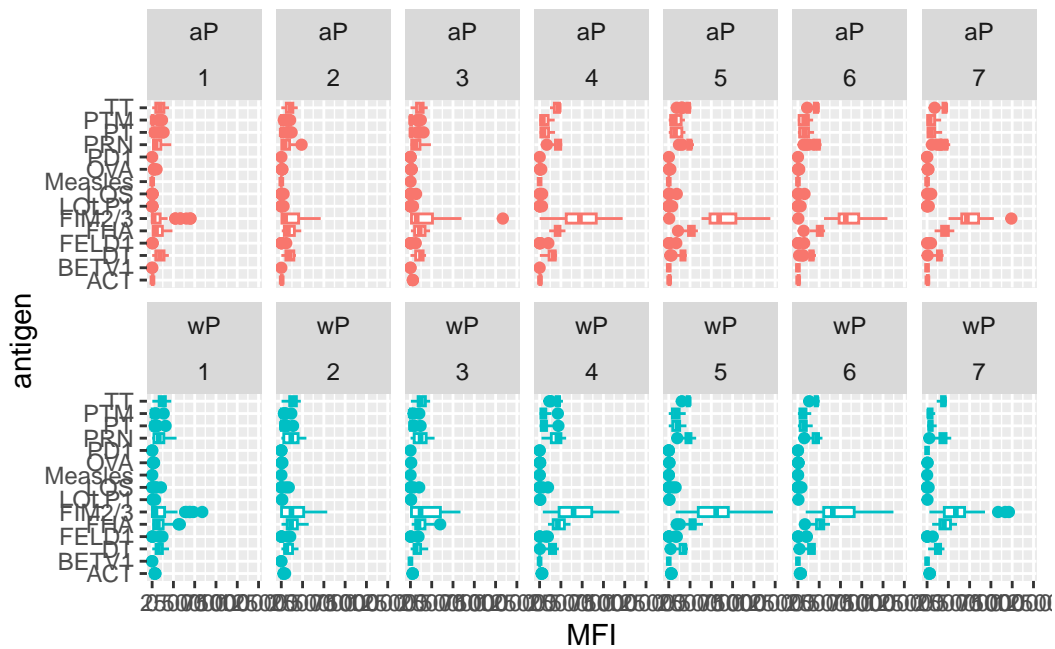
adding infancy\_vac to the faceting

```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```



Another version of this plot adding infancy\_vac to the faceting:

```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```



```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```

