Class 15 - Pertusssis Mini Project

AUTHOR

Achyuta

Pertussis data by year:

CDC Data

We will use the datapasta R package to “scrape” this data into R.

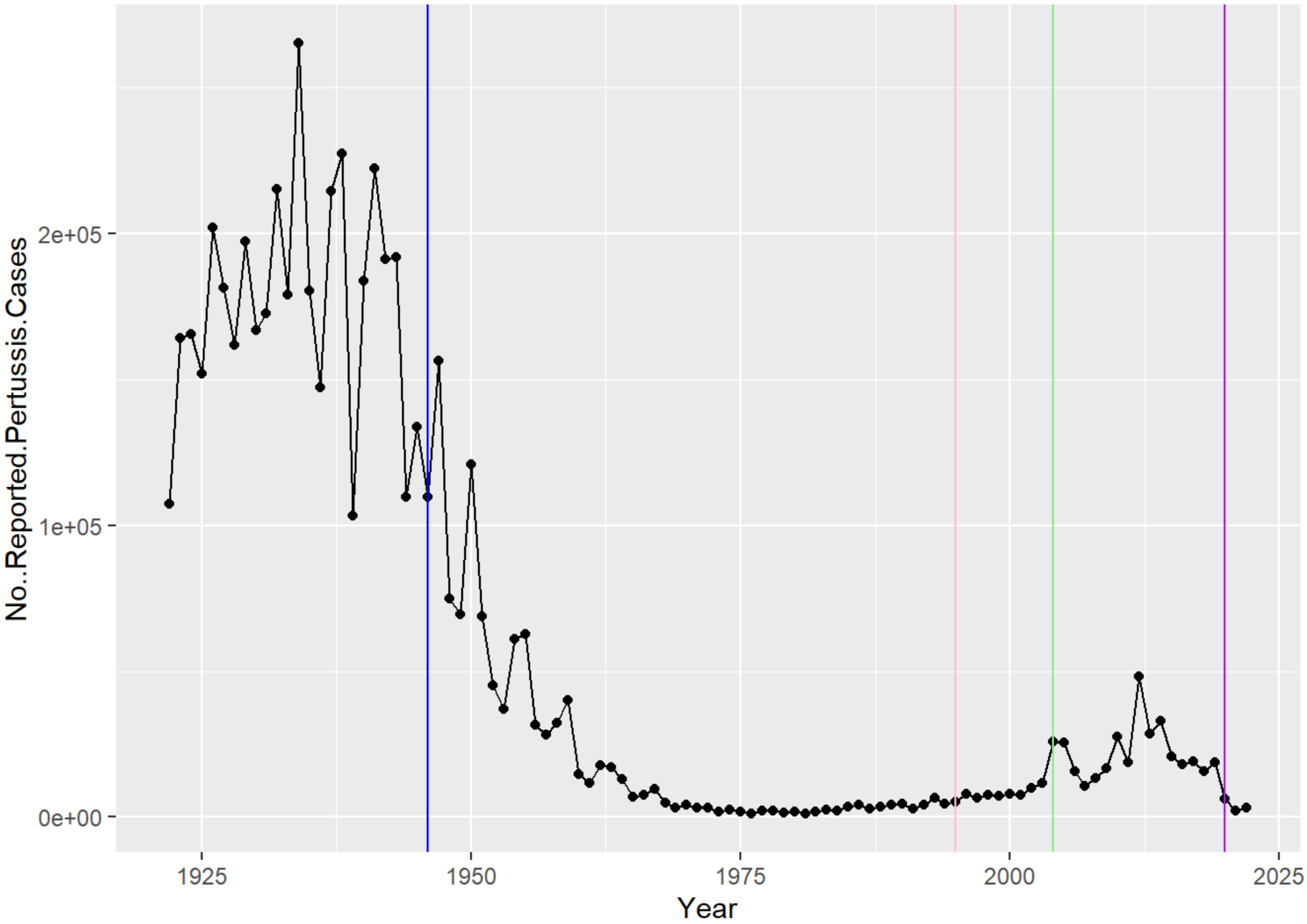
```
cdc <- data.frame(
                                 Year = c(1922L,1923L,1924L,1925L,
                                          1926L,1927L,1928L,1929L,1930L,1931L,
                                          1932L,1933L,1934L,1935L,1936L,
                                          1937L,1938L,1939L,1940L,1941L,1942L,
                                          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),
         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)
       )
```

```
library(ggplot2)
baseplot <- ggplot(cdc) +
  aes(x = Year, y = No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line()
```

Add some landmark developments as annotation to our plot. We include the first whole-cell (wP) vaccine roll-out in 1940.

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

```
baseplot +
  geom_vline(xintercept = 1946, col = "blue") +
  geom_vline(xintercept = 1995, col = "pink") +
  geom_vline(xintercept = 2020, col = "purple") +
  geom_vline(xintercept = 2004, col = "light green")
```



We went from ~200,000 cases pre wP vaccine to ~1,000 cases in 1976. The US switched to the aP vaccine in 1995. We start to see a big increase in 2004 to ~26,000 cases.

There is a ~10 year lag from aP roll-out to increasing case numbers. This holds true of other countries like Japan, UK, etc.

Key queestion: Why does the aP vaccine induced immunity wane faster than that of the wP vaccine?

##CMI-PB

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

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

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

Let's read the first one of these tables.

```
library(jsonlite)

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
```

> Q1. How many subjects are there in this datase?

```
nrow(subject)
```

```
[1] 172
```

> Q2. How many aP and wP individuals are there?

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

```
aP wP
87 85
```

> Q3. How many males and females are there?

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

```

Female   Male
   112     60
```

> Q4. Breakdown by biological sex and race.

```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
```

> Q Does this do a good job of representing the US populus?

No

Let's get more data from CMI-PB, this time about the specimens collected.

```r
specimen <- read_json("https://www.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
```

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

```
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                            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("https://www.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 join to do of `meta` and `abdata` to associate all the metadata about the individual and their race, biological sex, and infancy vaccination status together with antibody levels

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

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

```
head(ab)
```

```
  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
```

```r
nrow(ab)
```

```
[1] 52576
```

How many isotypes

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

```

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

How many antigens?

```r
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 
```

Let’s focus in on IgG - one of the main antibody types responsive to bacterial or viral infections.

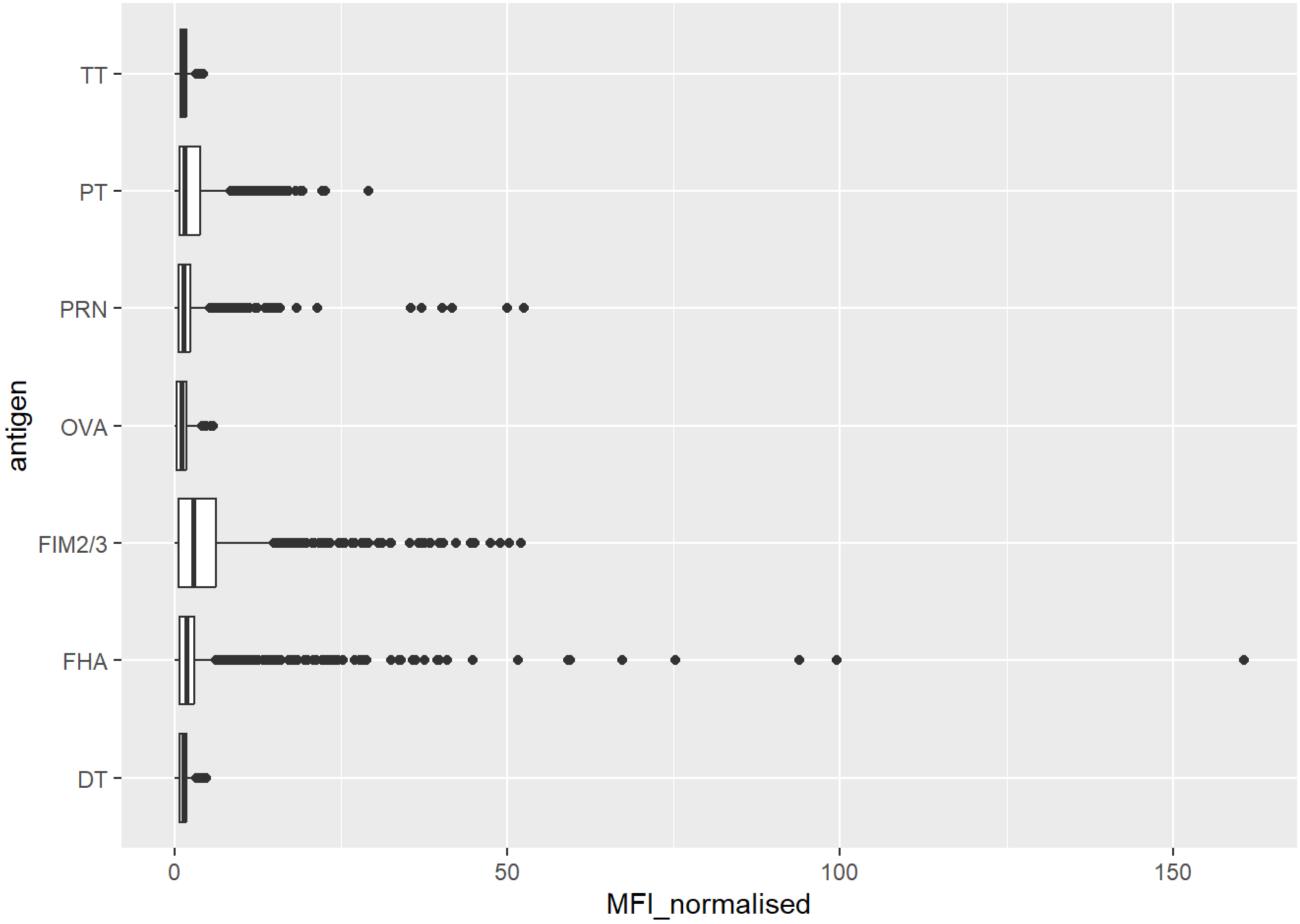
```r
igg <- filter(ab, isotype == "IgG")
```

```r
head(igg)
```

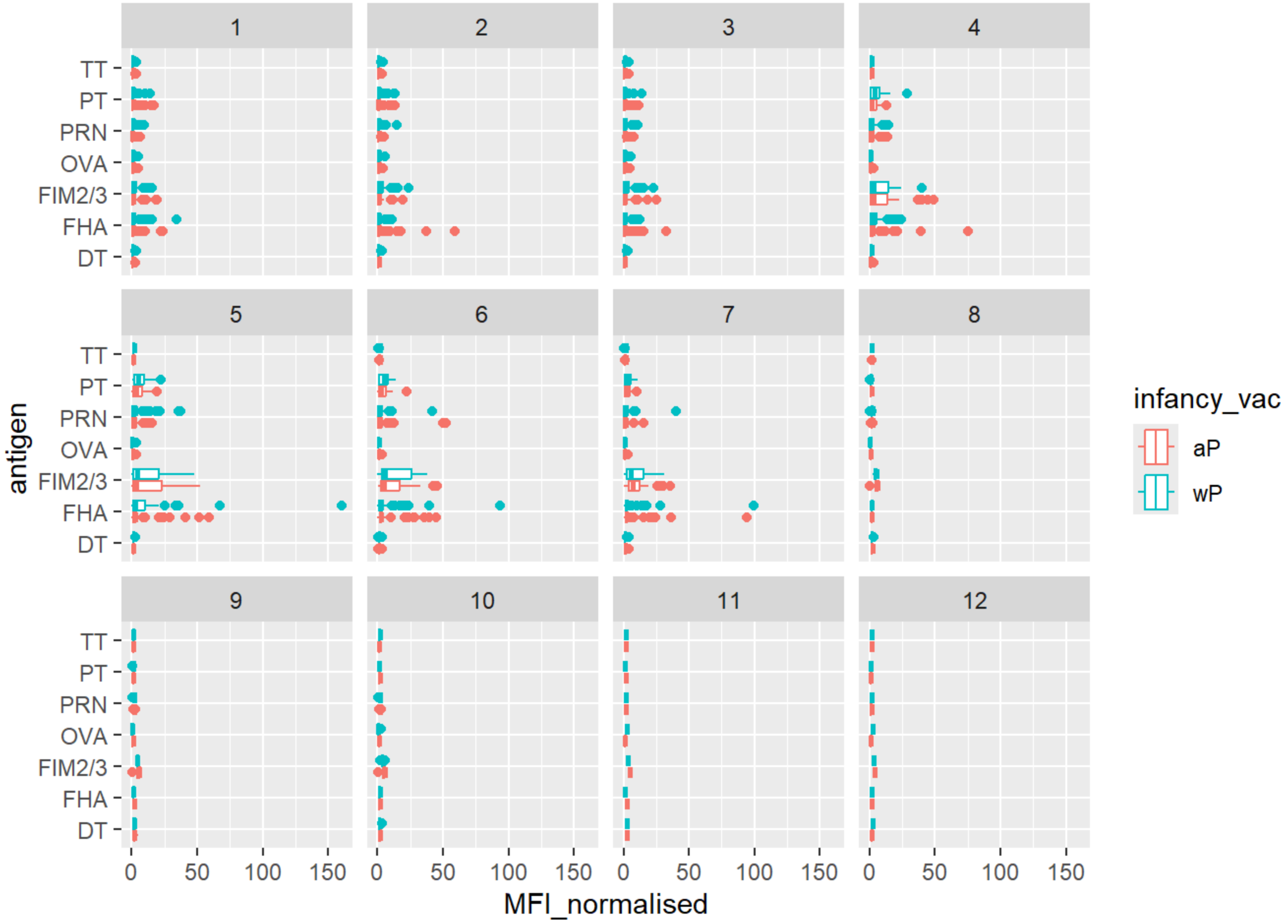
```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
3          1          wP         Female Not Hispanic or Latino White
4          1          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
```

Make a first plot of MFI (Mean Fluorescence Intensity - measure of how much is detected) for each antigen.

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



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



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

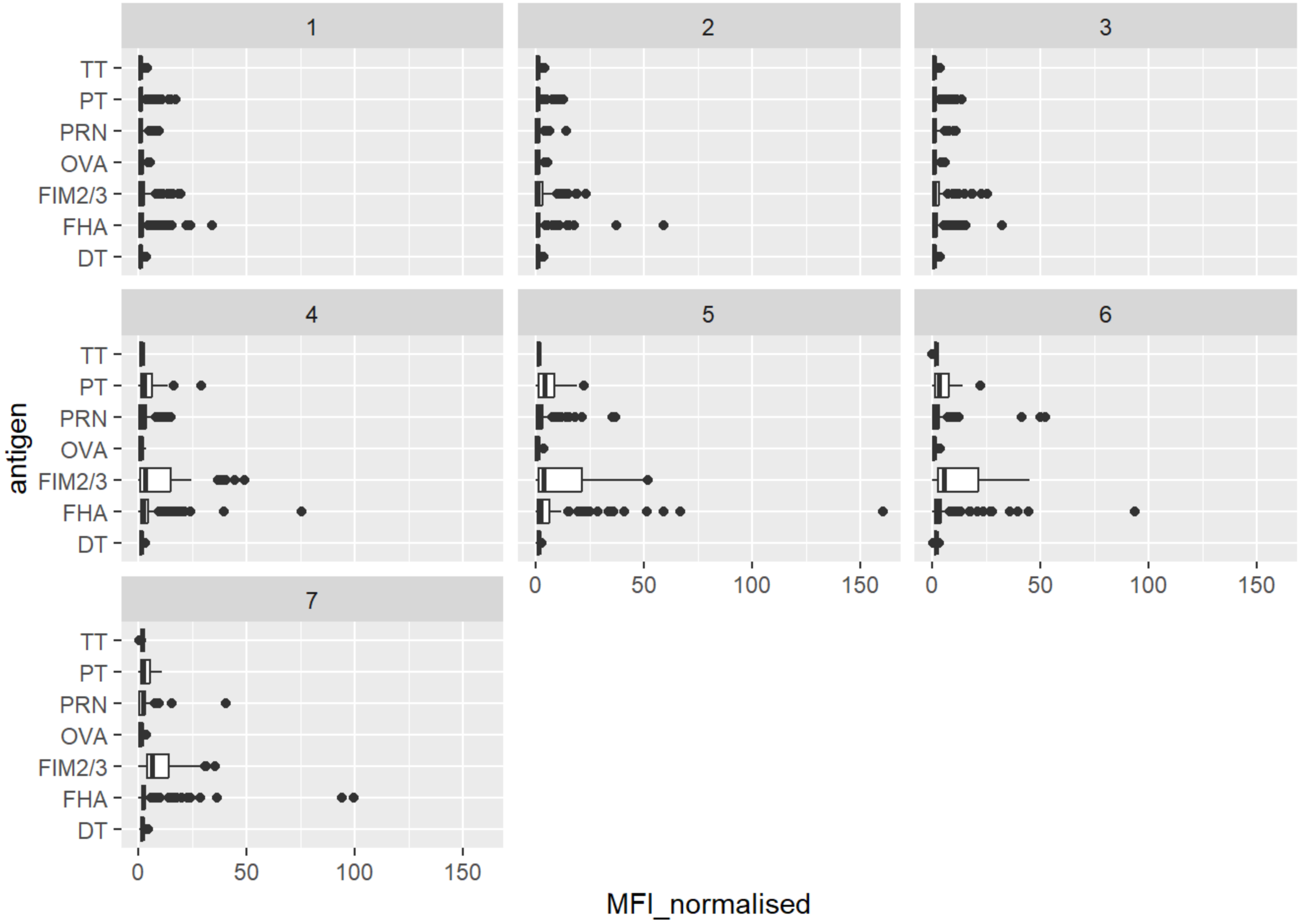
```

  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 let’s exclude these.

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

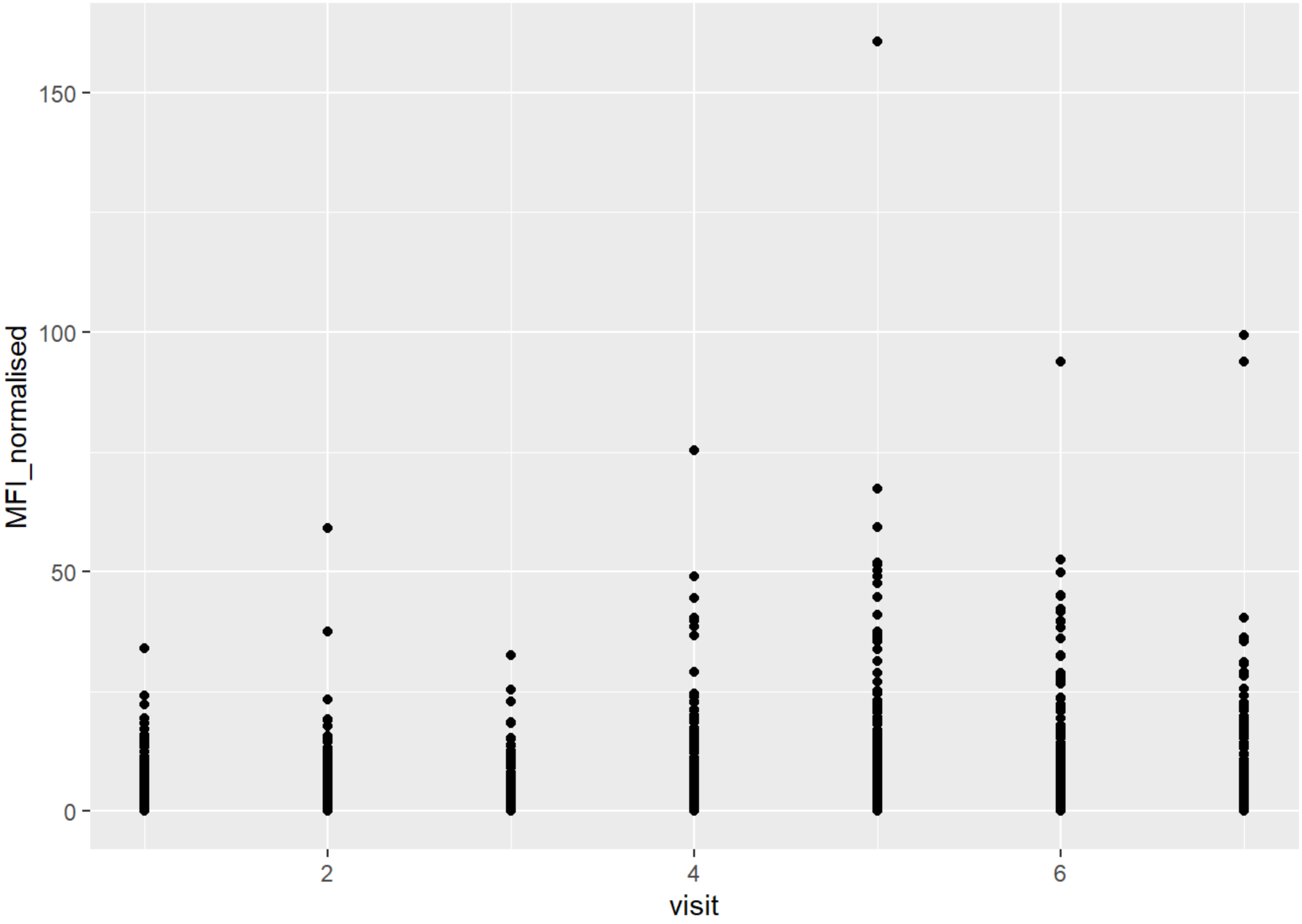
```

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

```
ggplot(igg_7) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
  facet_wrap(~visit)
```

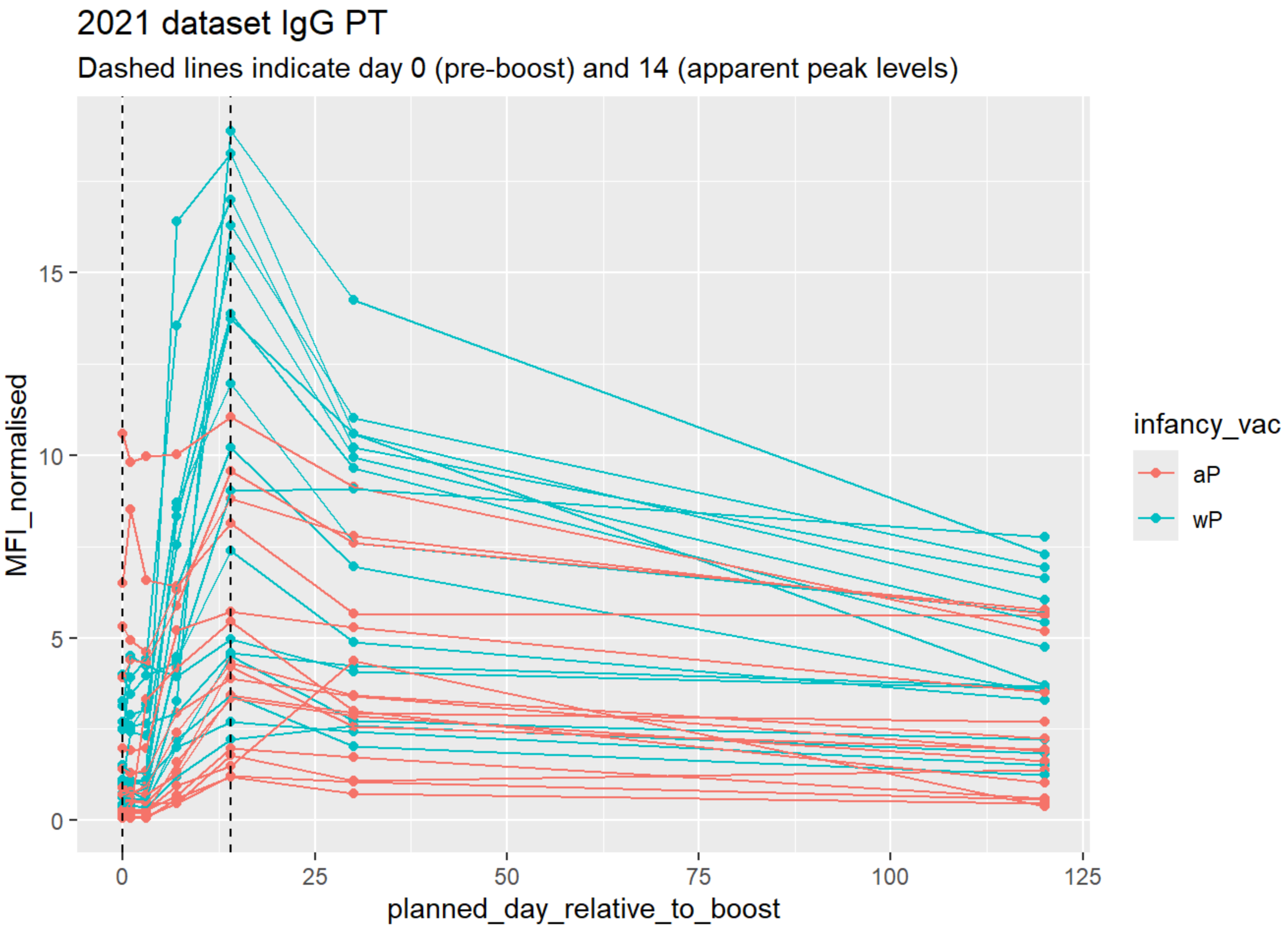


Let’s try a different plot. First focus on one antigen, start with PT (Pertussis Toxin) and plot visit or time on the x-axis and MFI normalized on the y axis.

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



```
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)")
```



2021 dataset IgG PT

Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)

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

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

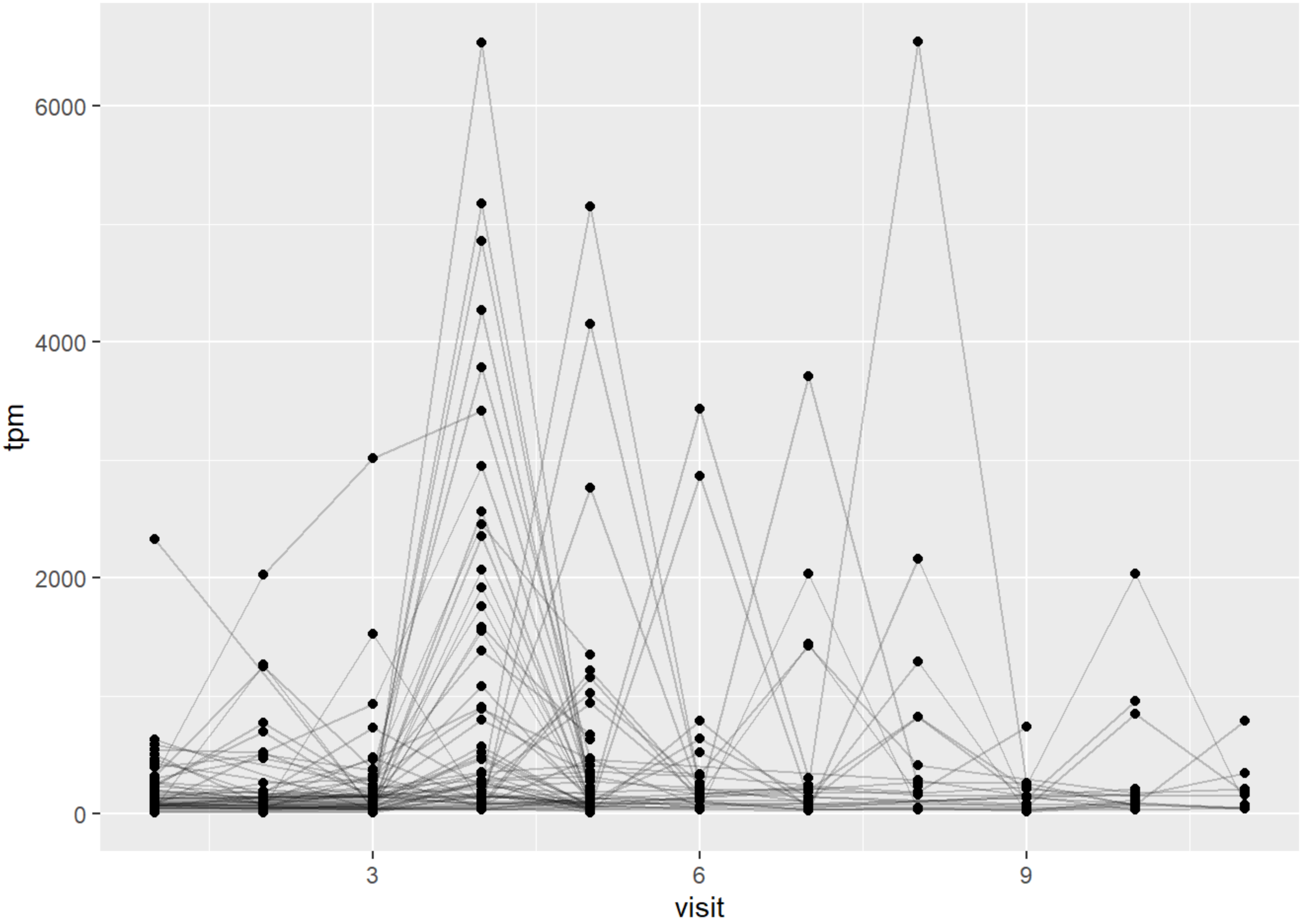
```
meta <- inner_join(specimen, subject)
```

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

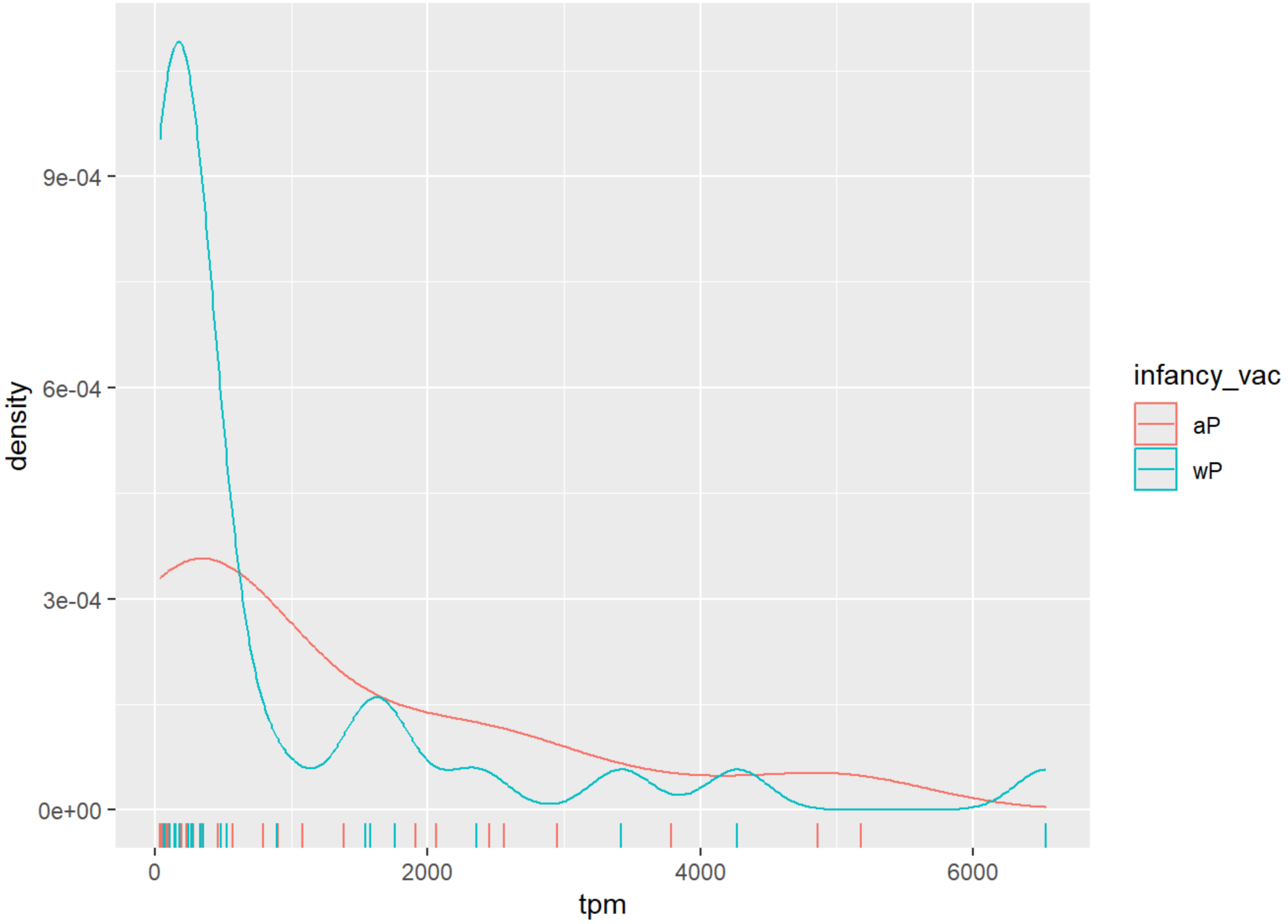
```
ssrna <- inner_join(rna, meta)
```

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

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



```
ssrna %>%
  filter(visit==4) %>%
  ggplot() +
    aes(tpm, col=infancy_vac) + geom_density() +
    geom_rug()
```


9e-04
6e-04
3e-04
0e+00
density
0
2000
4000
6000
tpm
infancy_vac
aP
wP