Class 18: Pertussis Mini Project

Ruth Barnes: A16747659

2025-02-06

Table of contents

Pertussis and the CMI-PB project

Background

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

1. Investigating pertussis cases by year

The CDC tracks Pertussis cases around the US: http://tinyurl.com/pertussiscdc

We can "scrape" this data using the R **datapasta** package:

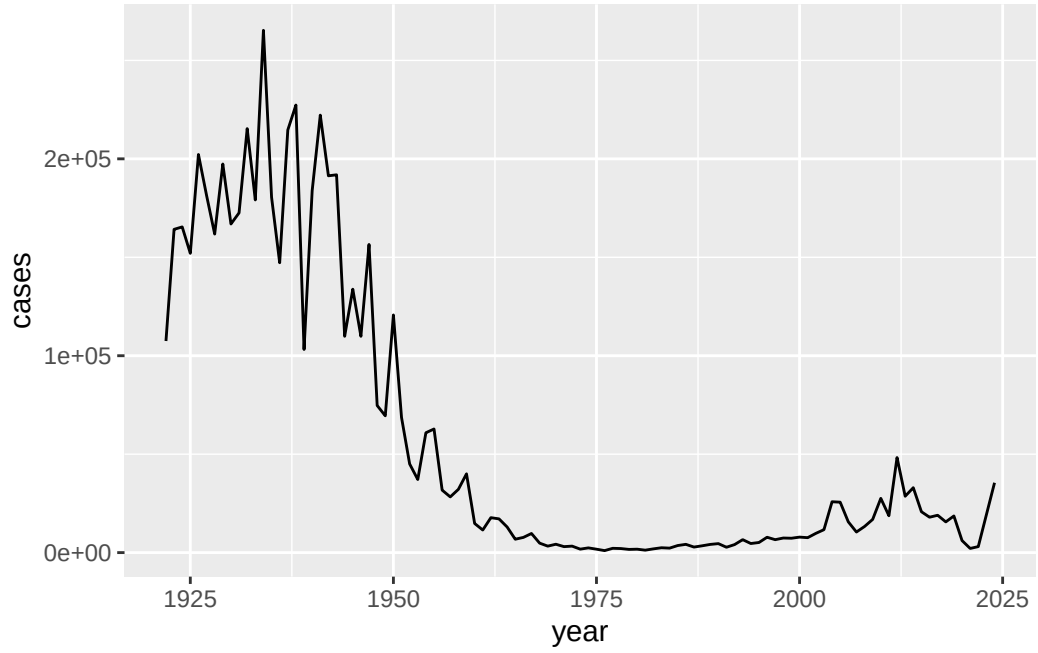
Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
head(cdc)
```

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

```
library(ggplot2)

ggplot(cdc) +
  aes(year, cases) +
  geom_line()
```



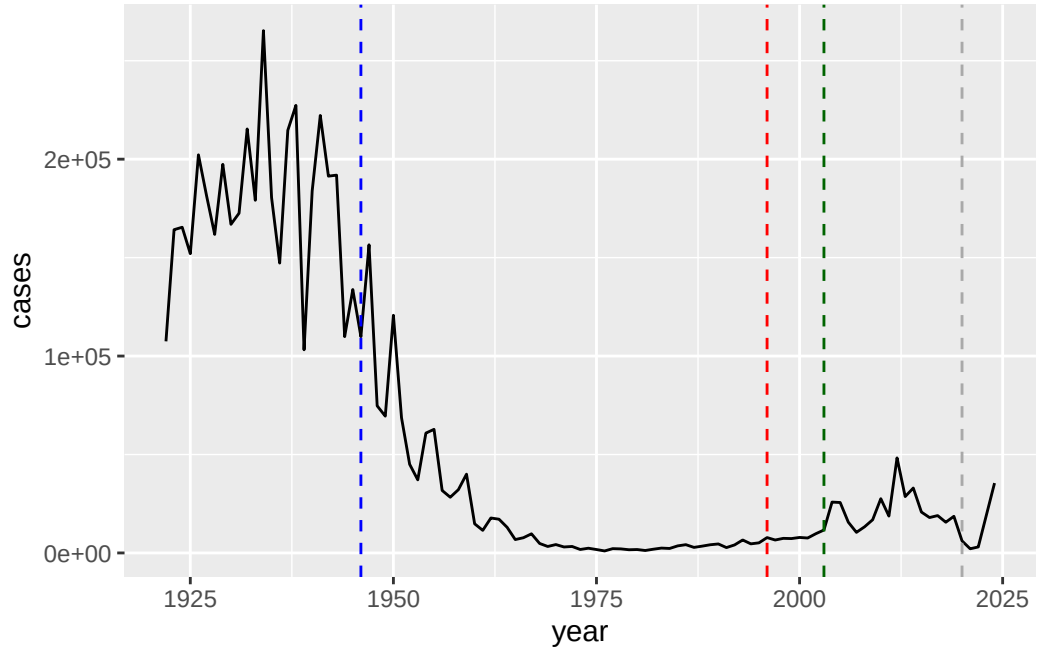
2. A tale of two vaccines (wP & aP)

Two types of pertussis vaccines have been developed: whole-cell pertussis (wP) and acellular pertussis (aP). The first vaccines were composed of 'whole cell' (wP) inactivated bacteria. The latter aP vaccines use purified antigens of the bacteria.

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?

```
library(ggplot2)

ggplot(cdc) +
  aes(year, cases) +
  geom_line() +
  geom_vline(xintercept = 1946, color = "blue", linetype = "dashed") +
  geom_vline(xintercept = 1996, color = "red", linetype = "dashed") +
  geom_vline(xintercept = 2020, color = "darkgrey", linetype = "dashed") +
  geom_vline(xintercept = 2003, color = "darkgreen", linetype = "dashed")
```



There were high case 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 fo Pertussis again. There is also a notable COVID realted 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?

After the introduction of the aP vaccine, there is a peak in case numbers due to waning immunity and the aging of the vaccine (need booster shots).

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

The CMI-PB API returns JSON data

The CMI-PB project aims to address this key question: What is different between the immune response to infection if you had an older wP vaccine vs the newer aP vaccine?

We can look at all the data from this ongoing project via JSON API calls. For this we will use the **jsonlite** package. We can install with: `install.packages("jsonlite")`

```
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
```

Q. How many individials "subjects" are in this data set?

```
nrow(subject)
```

```
[1] 172
```

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

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

```
aP wP
87 85
```

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

```
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...)?

```
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
```

This is not representative of the US population but it is the biggest data-set of its type so let's see what we can learn...

Side-Note: Working with dates

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?

```
library(lubridate)
```

```
Attaching package: 'lubridate'
```

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

    date, intersect, setdiff, union
```

```r
# Calculate age in years using a fixed date
subject$age <- time_length(today("2025-03-08") - ymd(subject$year_of_birth), "years")
```

```
Warning in with_tz.default(Sys.time(), tzone): Unrecognized time zone
'2025-03-08'
```

```
Warning in as.POSIXlt.POSIXct(x, tz = tz): unknown timezone '2025-03-08'
```

```r
print(subject$age)
```

```
  [1] 39.18686 57.18823 42.18754 37.18823 34.18754 37.18823 44.18617 40.18617
  [9] 29.18823 43.18686 39.18686 43.18686 28.18617 32.18617 36.18617 38.18754
 [17] 45.18823 28.18617 31.18686 44.18617 42.18754 40.18617 34.18754 33.18823
 [25] 37.18823 42.18754 28.18617 43.18686 28.18617 37.18823 36.18617 28.18617
 [33] 35.18686 42.18754 34.18754 28.18617 27.18686 28.18617 40.18617 31.18686
 [41] 40.18617 28.18617 27.18686 27.18686 28.18617 27.18686 29.18823 27.18686
 [49] 28.18617 28.18617 28.18617 27.18686 27.18686 28.18617 28.18617 28.18617
 [57] 29.18823 28.18617 28.18617 28.18617 38.18754 32.18617 30.18754 32.18617
 [65] 35.18686 49.18823 53.18823 53.18823 35.18686 27.18686 27.18686 34.18754
 [73] 30.18754 30.18754 27.18686 27.18686 37.18823 32.18617 38.18754 33.18823
 [81] 32.18617 27.18686 26.18754 28.18617 25.18823 27.18686 25.18823 25.18823
 [89] 28.18617 26.18754 27.18686 25.18823 29.18823 26.18754 27.18686 25.18823
 [97] 39.18686 32.18617 26.18754 24.18617 22.18754 22.18754 31.18686 36.18617
[105] 31.18686 29.18823 27.18686 30.18754 36.18617 28.18617 29.18823 29.18823
[113] 29.18823 35.18686 23.18686 25.18823 31.18686 27.18686 27.18686 30.18754
[121] 25.18823 26.18754 29.18823 25.18823 32.18617 32.18617 29.18823 31.18686
[129] 34.18754 29.18823 27.18686 30.18754 28.18617 35.18686 30.18754 30.18754
[137] 27.18686 25.18823 32.18617 24.18617 29.18823 34.18754 22.18754 26.18754
[145] 23.18686 33.18823 25.18823 37.18823 34.18754 34.18754 33.18823 30.18754
[153] 27.18686 28.18617 28.18617 24.18617 28.18617 25.18823 31.18686 29.18823
[161] 32.18617 26.18754 32.18617 34.18754 32.18617 24.18617 28.18617 34.18754
[169] 22.18754 33.18823 22.18754 39.18686
```

```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
# (i) Average age of aP individuals

wp <- subject %>% filter(infancy_vac == "wP")
avg_wp <- mean(wp$age, na.rm = TRUE)
print(avg_wp)
```

```
[1] 35.83428
```

```r
# (ii) Average age of aP individuals
ap <- subject %>% filter(infancy_vac == "aP")
avg_ap <- mean(ap$age, na.rm = TRUE)
print(avg_ap)
```

```
[1] 27.08358
```

```r
# (iii) Statistical significance test (t-test)
t_test <- t.test(wp$age, ap$age, var.equal = TRUE)

# Print results
cat("Average age of wP individuals:", avg_wp, "\n")
```

```
Average age of wP individuals: 35.83428
```

```r
cat("Average age of aP individuals:", avg_ap, "\n")
```

```
Average age of aP individuals: 27.08358
```

```r
print(t_test)
```

```
    Two Sample t-test

data:  wp$age and ap$age
t = 13.036, df = 170, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.425601 10.075807
sample estimates:
mean of x mean of y
 35.83428  27.08358
```

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

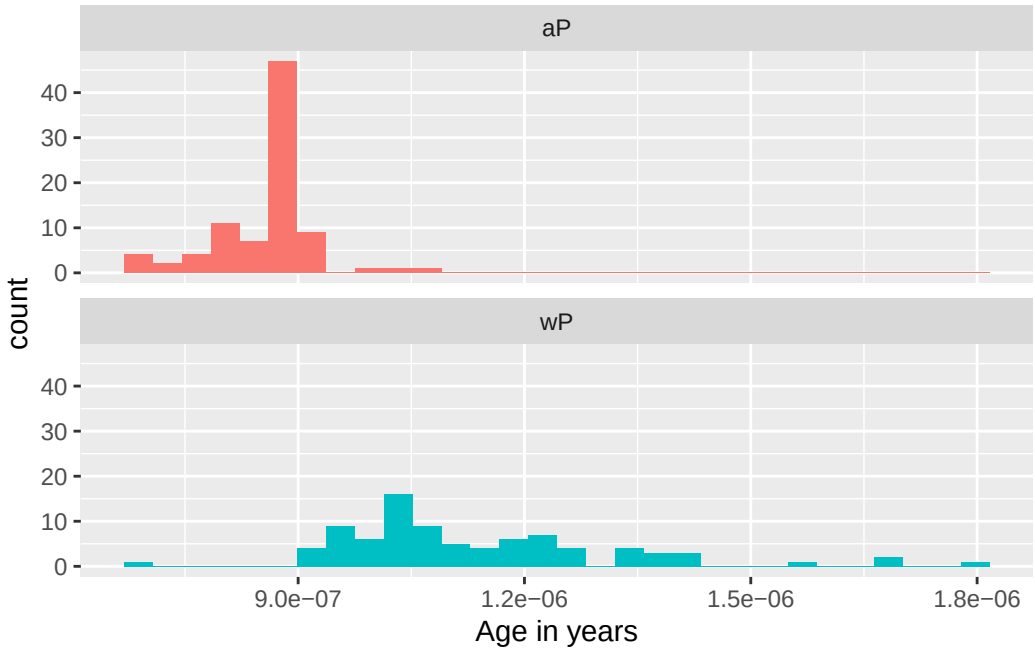
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

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

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Joining multiple tables

Obtain more data from CMI-PB:

```
specimen <- read_json("http://cmi-pb.org/api/v5_1/specimen", simplifyVector = T)
ab_data <- read_json("http://cmi-pb.org/api/v5_1/plasma_ab_titer", simplifyVector = T)
```

```
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
```

```
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
```

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 u need to work with.

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

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:

```
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 39.18686           1
2    1986-01-01    2016-09-12 2020_dataset 39.18686           2
3    1986-01-01    2016-09-12 2020_dataset 39.18686           3
4    1986-01-01    2016-09-12 2020_dataset 39.18686           4
5    1986-01-01    2016-09-12 2020_dataset 39.18686           5
6    1986-01-01    2016-09-12 2020_dataset 39.18686           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   9
```

```
dim(specimen)
```

```
[1] 1503    6
```

```
dim(meta)
```

```
[1] 1503   14
```

Now we can join our `ab_data` table to `meta` so we have all the information we need about 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.

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

```
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      age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 39.18686           1
2    1986-01-01    2016-09-12 2020_dataset 39.18686           1
3    1986-01-01    2016-09-12 2020_dataset 39.18686           1
4    1986-01-01    2016-09-12 2020_dataset 39.18686           1
5    1986-01-01    2016-09-12 2020_dataset 39.18686           1
6    1986-01-01    2016-09-12 2020_dataset 39.18686           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 different antibody isotypes are there in this dataset?/Q11. How many specimens (i.e. entries in abdata) do we have for each isotype?

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

```
[1] 61956
```

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

```
  IgE   IgG  IgG1  IgG2  IgG3  IgG4
 6698  7265 11993 12000 12000 12000
```

```
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
```

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

```
unique(abdata$dataset)
```

```
[1] "2020_dataset" "2021_dataset" "2022_dataset" "2023_dataset"
```

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

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

The most recent dataset, 2023_dataset, has 15,050 rows, which is an increase from 2021 and 2022 but still less than half of 2020's total

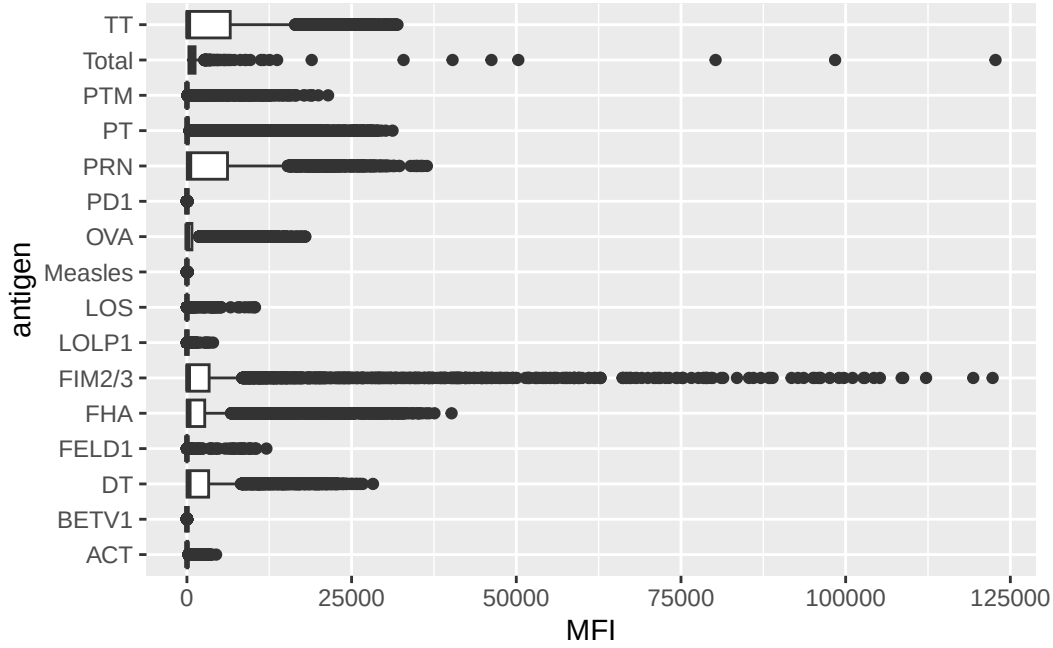
I want a plot of antigen levels, across the whole dataset.

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

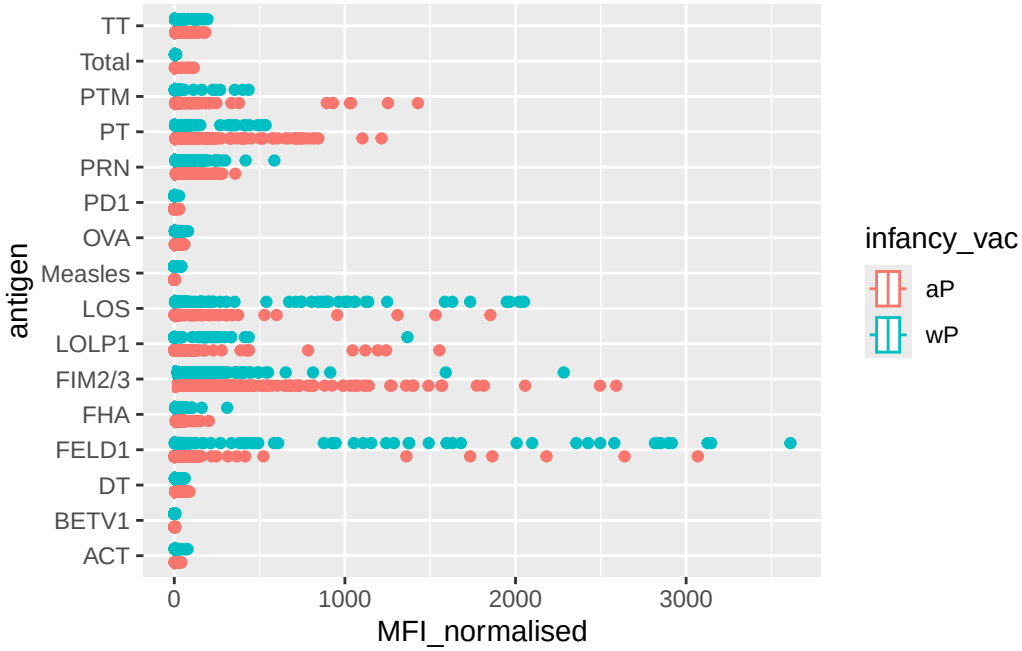
```
library(ggplot2)

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

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



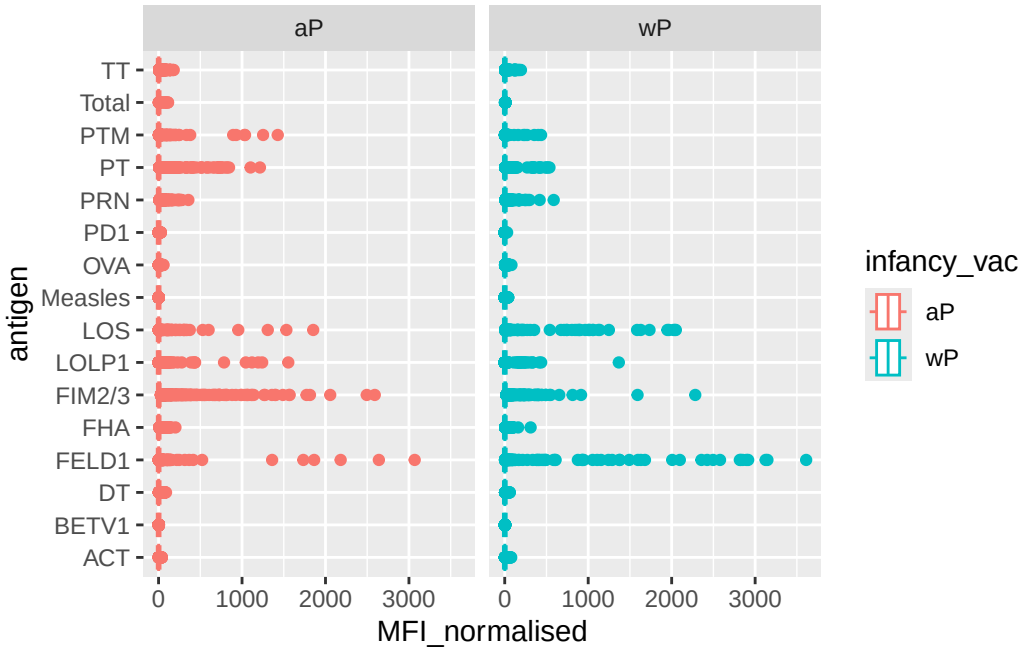

```
ggplot(abdata) +
  aes(MFI_normalised, antigen, col=infancy_vac) +
  geom_boxplot()
```



Antigens like FIM2/3, PT, FELD1 have quite a large range of values. Others like Measles don't show much activity.

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

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



Anitgens FIM2/3, LOS, and FELD1 show slightly similar levels of IgG antibody titers. These antigen levels are different due to variations in immune response, antigen exposure, and immune memory.

4. Examine IgG Ab titer levels

For this I need to select out just iostype IgG.

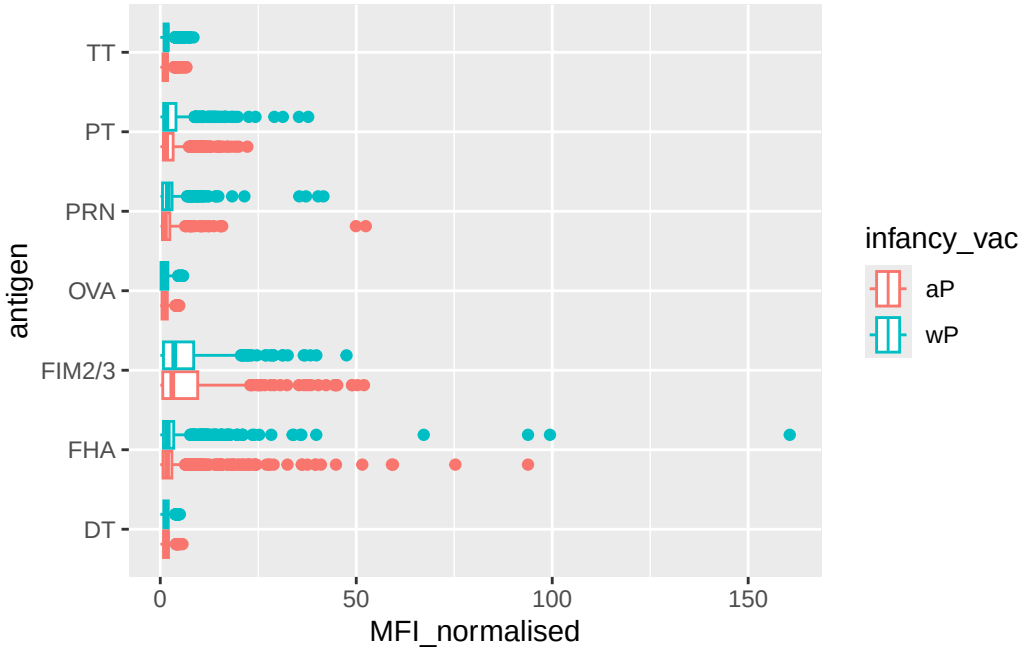
```
igg <- abdata |>
  filter(isotype == "IgG")
head(igg)
```

```
  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      age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 39.18686           1
2    1986-01-01    2016-09-12 2020_dataset 39.18686           1
```

```
3    1986-01-01    2016-09-12 2020_dataset 39.18686           1
4    1986-01-01    2016-09-12 2020_dataset 39.18686           2
5    1986-01-01    2016-09-12 2020_dataset 39.18686           2
6    1986-01-01    2016-09-12 2020_dataset 39.18686           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
```

A overveiw boxplot:

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

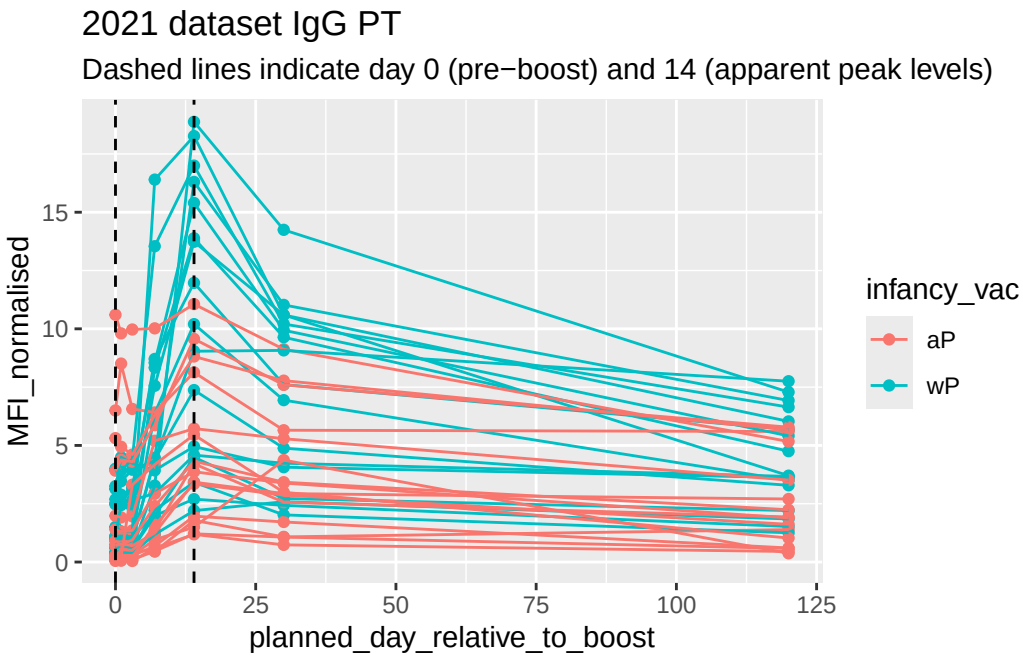


Digging in further to look at the time course of IgG isotype PT antigen levels actoss aP and wP individauls:

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

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

# Plot and color 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)")
```


2021 dataset IgG PT
Dashed lines indicate day 0 (pre−boost) and 14 (apparent peak levels)
MFI_normalised
15
10
5
0
0
25
50
75
100
125
planned_day_relative_to_boost
infancy_vac
aP
wP