Class 18 Investigating Pertussis Resurgence

AUTHOR
Erin McTavish PID: A17300519

Background

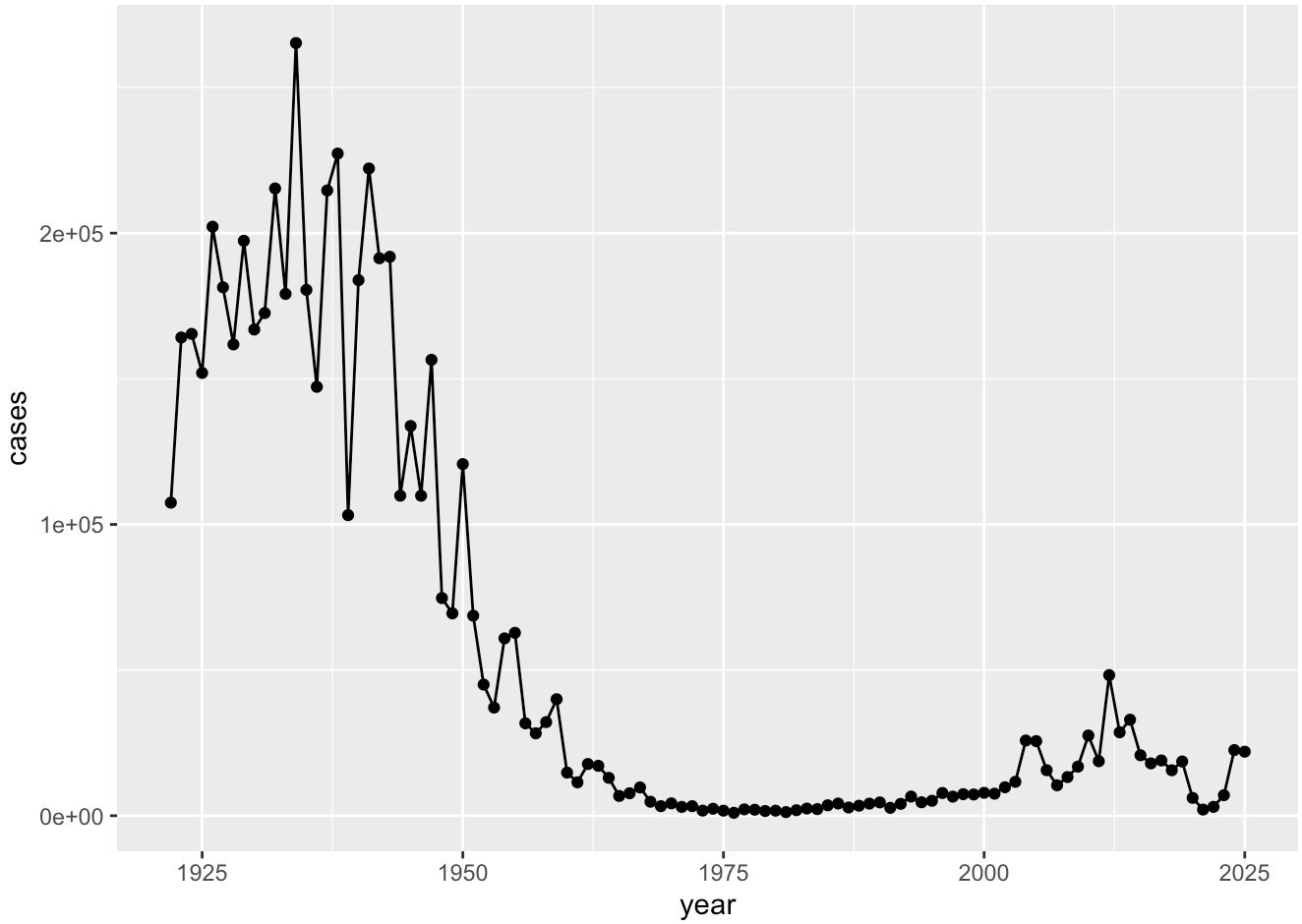
Pertussis (whopping cough) is a common lung infection caused by the bacteria B. Pertussis. It can infect anyone but is most deadly for infants (under 1 year old).

CDC tracking data

The CDC tracks the number of Pertussis cases: https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

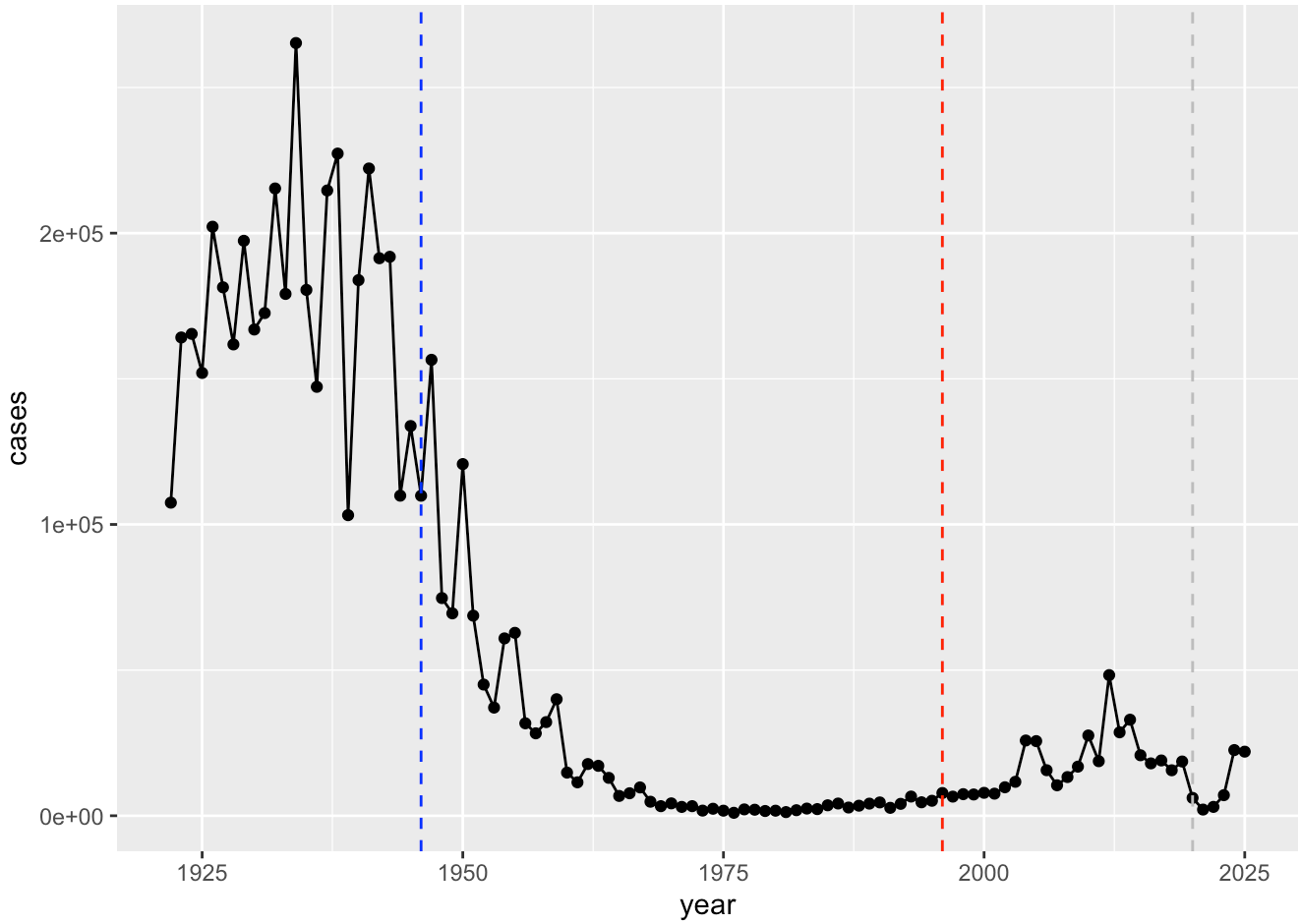
> Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
library(ggplot2)
ggplot(cdc) +
  aes(year, cases) +
  geom_point() +
  geom_line()
```



> 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?

```
library(ggplot2)
ggplot(cdc) +
  aes(year, cases) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept =1946, col="blue", lty=2) +
  geom_vline(xintercept = 1996, col="red", lty=2) +
  geom_vline(xintercept=2020, col= "gray", lty=2)
```



Cases are decreasing from 1946 to 1996. After the wP vaccine was introduced, causes decreased dramatically, plateauing in the 70s-80s.

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

When the switch to the aP vaccine occurs, things initially did not change much, but we are now seeing sparatic outbreaks in the early 2000s-present potentually due to immunity decreasing/pathogens adapting to the vaccine/less people are being vaccinated/the aP vaccine could be less effective. Moreover, there may be a waning efficacy of the aP vaccicne, as time passes, the vaccine becomes less effective in your body.

Exploring CMI-PB data

The CMI-PB project < https://www.cmi-pb.org/ > mission of CMI-PB is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of Pertussis booster vaccination.

Basically, make available a large dataset on the immune respomnse to Pertussis. They use a “booster” vaccination as a proxy for Pertussis infection.

They make their data available as JSON format API, we can read thie into R with the `read_json()` function from the **jsonlite** package:

```
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?

```
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
```

> Q. Is this representative of the US population?

Absolutely not. By race, there is a lack of Alaska Natives, African American/Black, and Native Hawaiian individuals.

We can now read mor etables from the CMI-PB database

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

```
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_titer)
```

```
  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
```

> 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
```

```
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
```

```
dob_col <- names(subject)[grepl("birth|dob", names(subject), ignore.case = TRUE)][1]

subject$age <- today() - ymd(as.character(subject[[dob_col]]))

ap <- subject %>% filter(infancy_vac == "aP")
round(summary(time_length(ap$age, "years")))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     23      27      28      28      29      35
```

```
wp <- subject %>% filter(infancy_vac == "wP")
round(summary(time_length(wp$age, "years")))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     23      33      35      37      40      58
```

> 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)
```

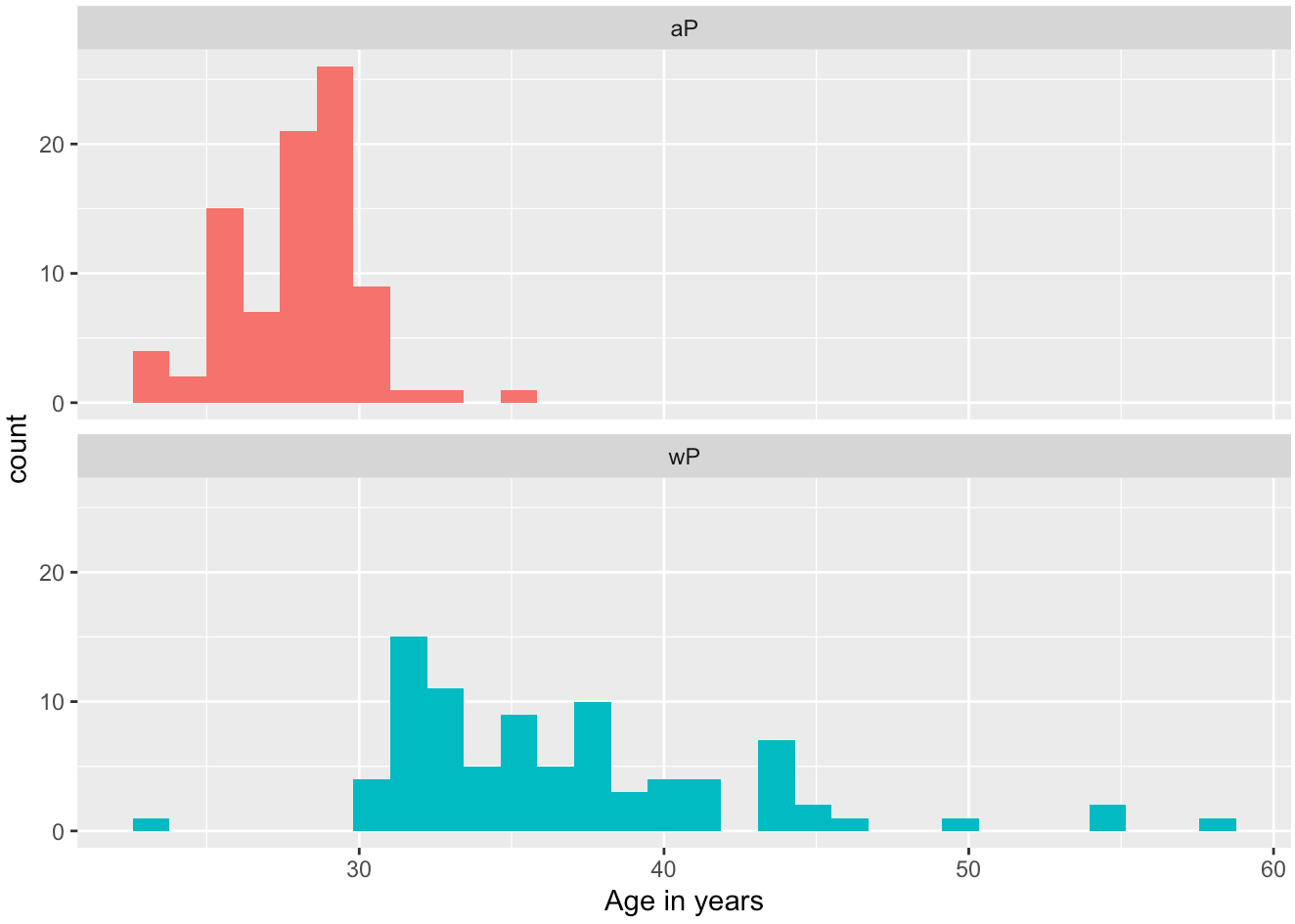
```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

To make sense of all this data we need to "join" (a.k.a. "merge" or "link") all these tables together. Only then will you know that a given Ab measurement (from the `ab_titer` table) was collected on a certain date (from the `speciment` table) from a certain wP or aP individual subject (from the `subjecy table).

> 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 `binwidth`.
```



```
# Or use wilcox.test()
x <- t.test(time_length( wp$age, "years" ),
       time_length( ap$age, "years" ))

x$p.value
```

```
[1] 2.372101e-23
```

We can use **dplyr** and the `*join()` family of functions to do this.

```
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 14673 days           1
2    1986-01-01    2016-09-12 2020_dataset 14673 days           2
3    1986-01-01    2016-09-12 2020_dataset 14673 days           3
4    1986-01-01    2016-09-12 2020_dataset 14673 days           4
5    1986-01-01    2016-09-12 2020_dataset 14673 days           5
6    1986-01-01    2016-09-12 2020_dataset 14673 days           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
```

Let's do one more `inner_join()` to join the `ab_titer` with all our `meta` data.

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

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

```
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 subject_id infancy_vac biological_sex
1 UG/ML                 2.096133          1          wP         Female
2 IU/ML                29.170000          1          wP         Female
3 IU/ML                 0.530000          1          wP         Female
4 IU/ML                 6.205949          1          wP         Female
5 IU/ML                 4.679535          1          wP         Female
6 IU/ML                 2.816431          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
        age actual_day_relative_to_boost planned_day_relative_to_boost
1 14673 days                          -3                             0
2 14673 days                          -3                             0
3 14673 days                          -3                             0
4 14673 days                          -3                             0
5 14673 days                          -3                             0
6 14673 days                          -3                             0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

> Q9 pt2. 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:

```
# Complete the API URLs...
specimen <- read_json("https://www.cmi-pb.org/api/v5_1/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/v5_1/plasma_ab_titer", simplifyVector = TR
```

```
# Join specimen and subject tables
meta <- left_join(specimen, subject)
```

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

```
dim(meta)
```

```
[1] 1503   14
```

```
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
         age
1 14673 days
2 14673 days
3 14673 days
4 14673 days
5 14673 days
6 14673 days
```

> 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(titer, meta)
```

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

```
dim(abdata)
```

```
[1] 61956    21
```

> 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  7265 11993 12000 12000 12000
```

> 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$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
```

Let’s focus on IgG isotype

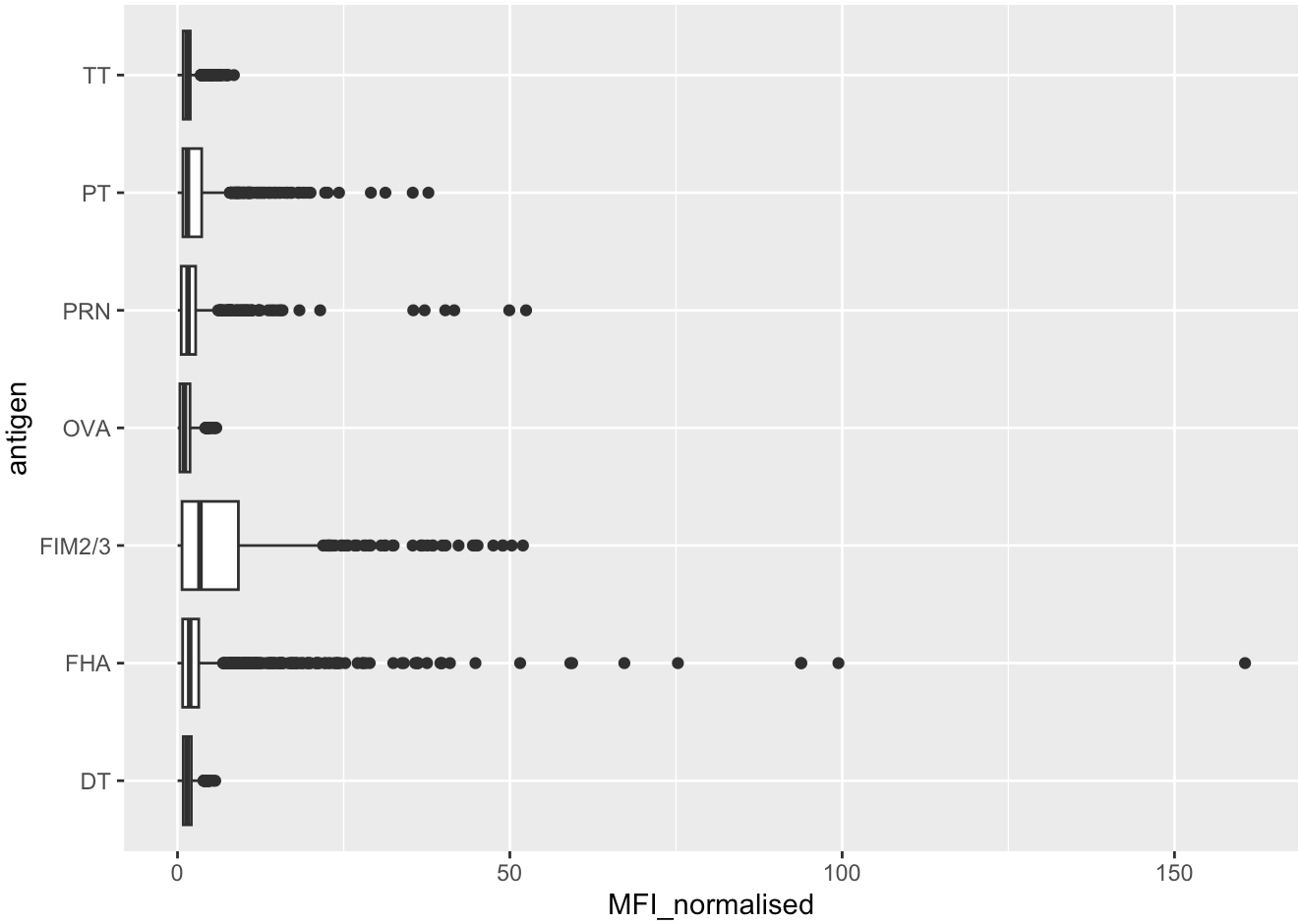
```
igg <- abdata |>
  filter(isotype=="IgG")

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
         age
1 14673 days
2 14673 days
3 14673 days
4 15769 days
5 15769 days
6 15769 days
```

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

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

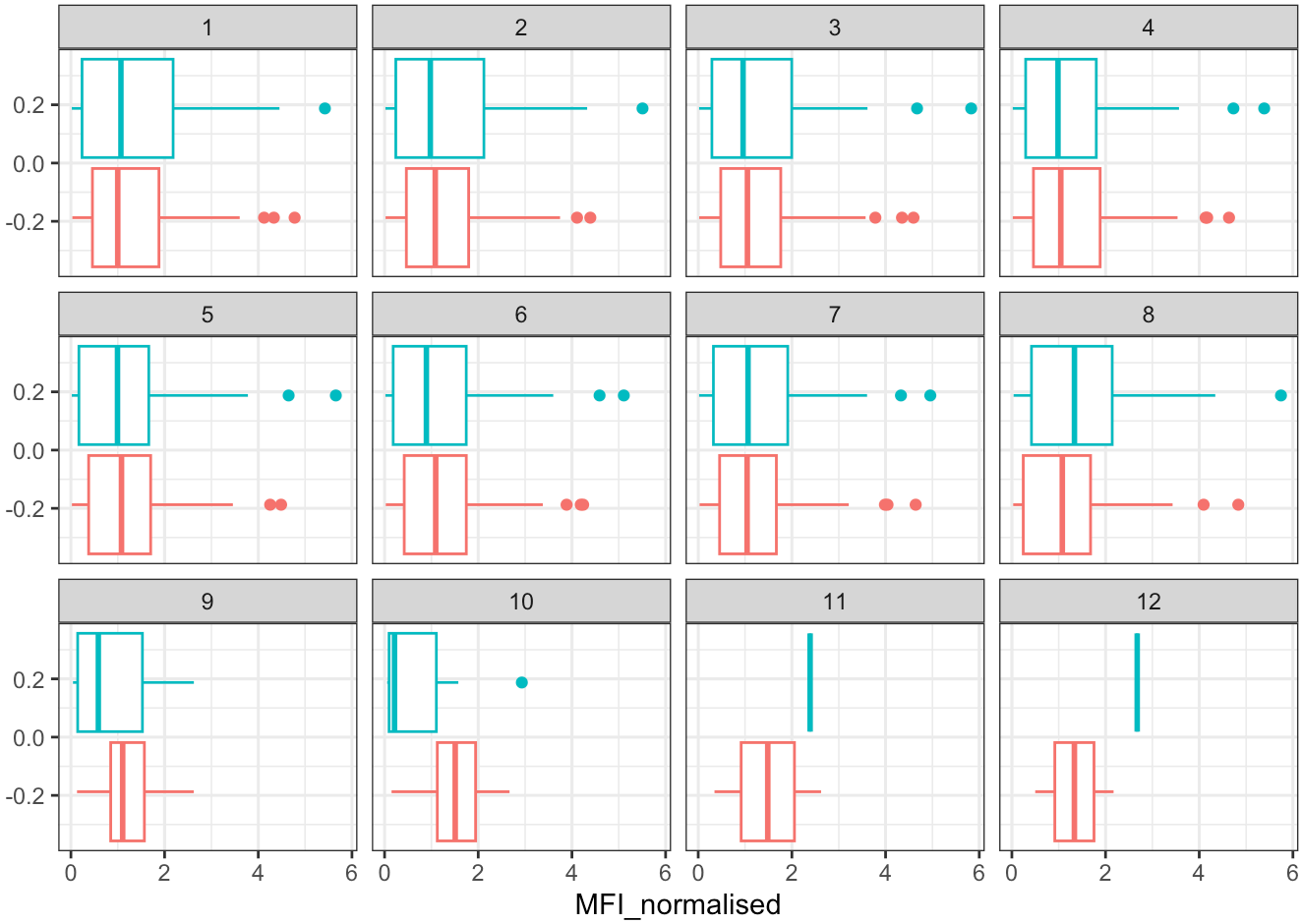


> Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

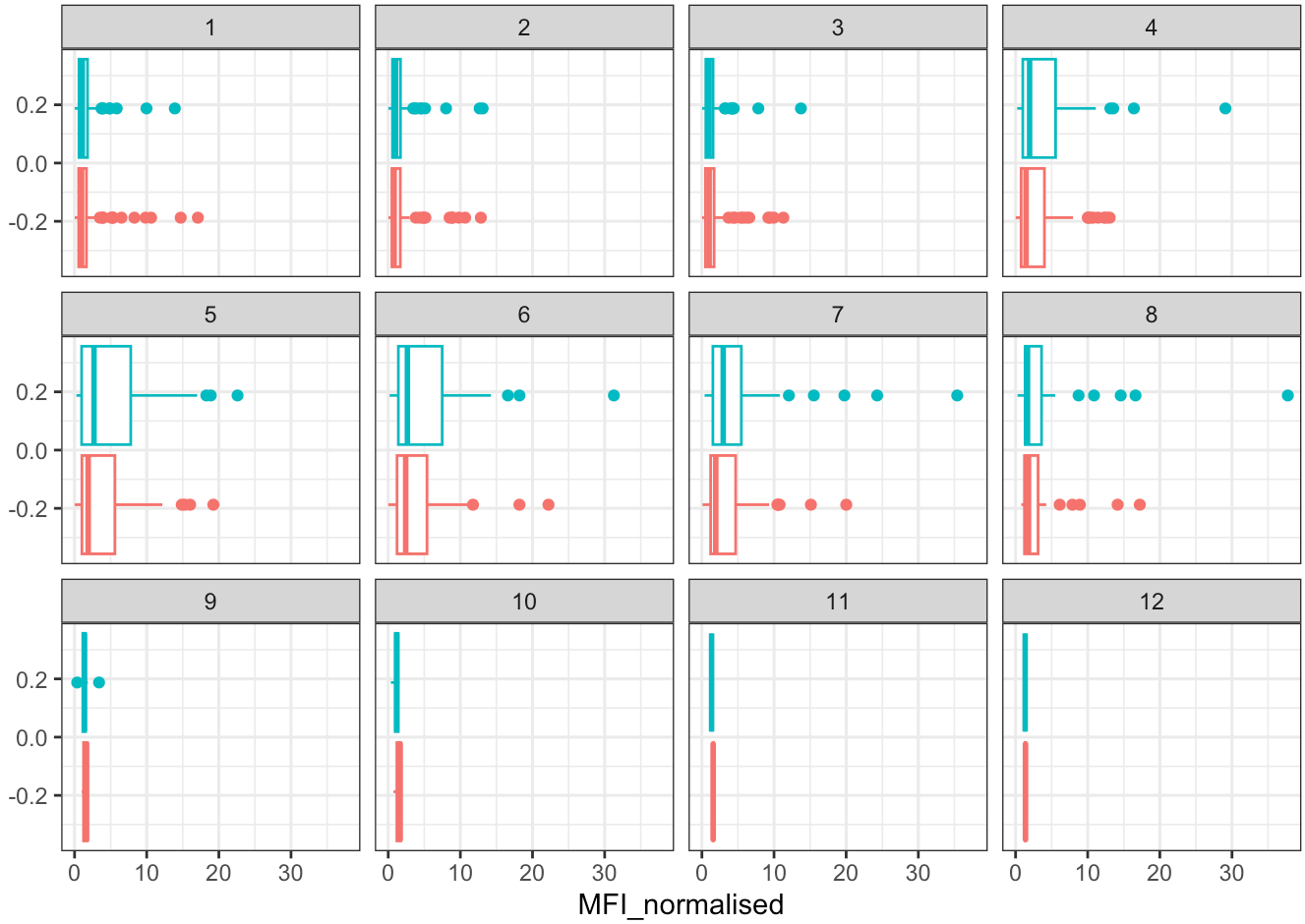
The antigens “PT”, “FIM2/3” and “FHA” appear to have the widest range of values. These are all components of the aP vaccine. These antigens have higher levels of IgG antitbody recognizing over time because you are being vaccinated by the aP booster shot and your immune system is recognizing and responding to it!

> 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).

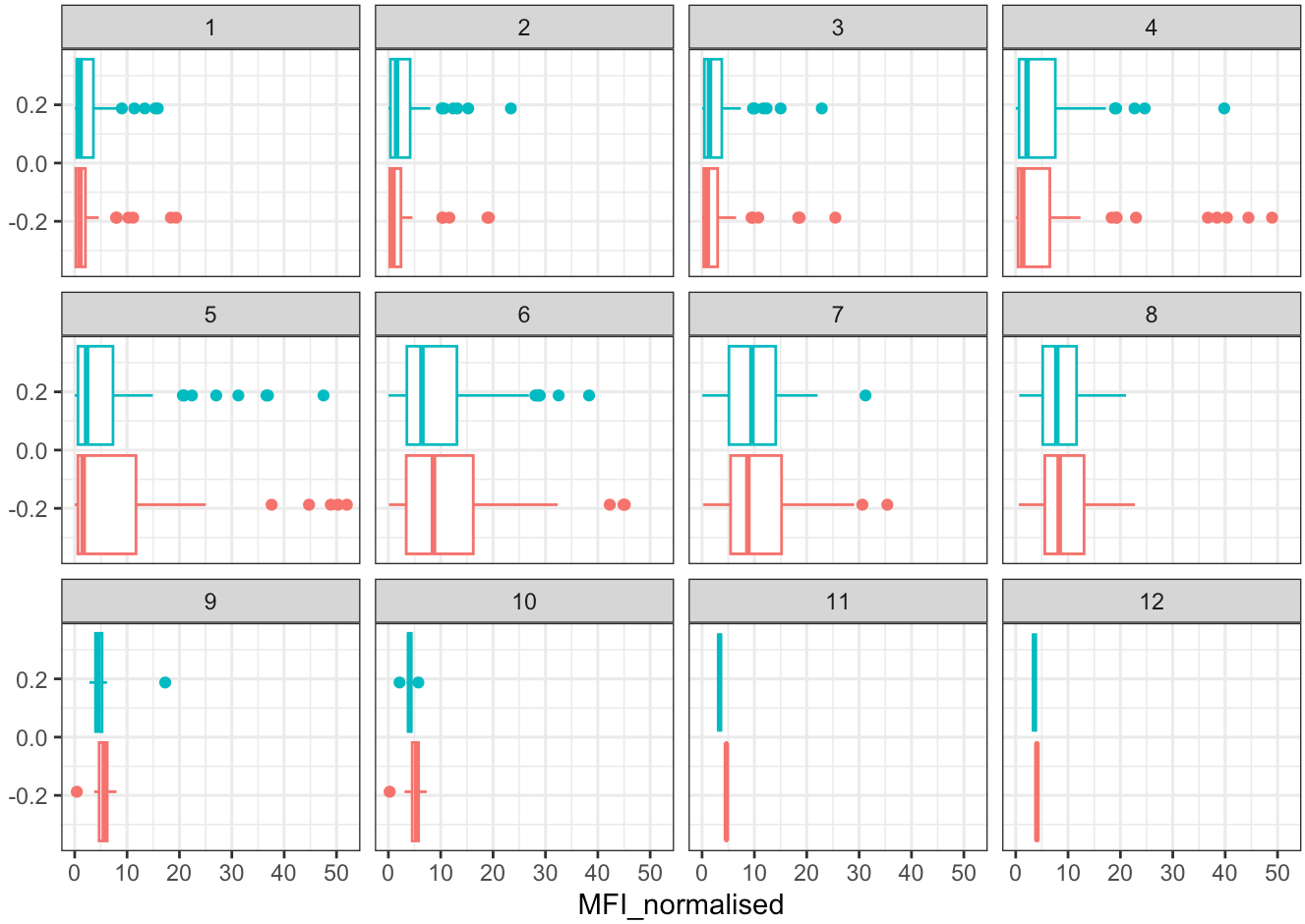
```
filter(igg, antigen=="OVA") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




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




```
filter(igg, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```

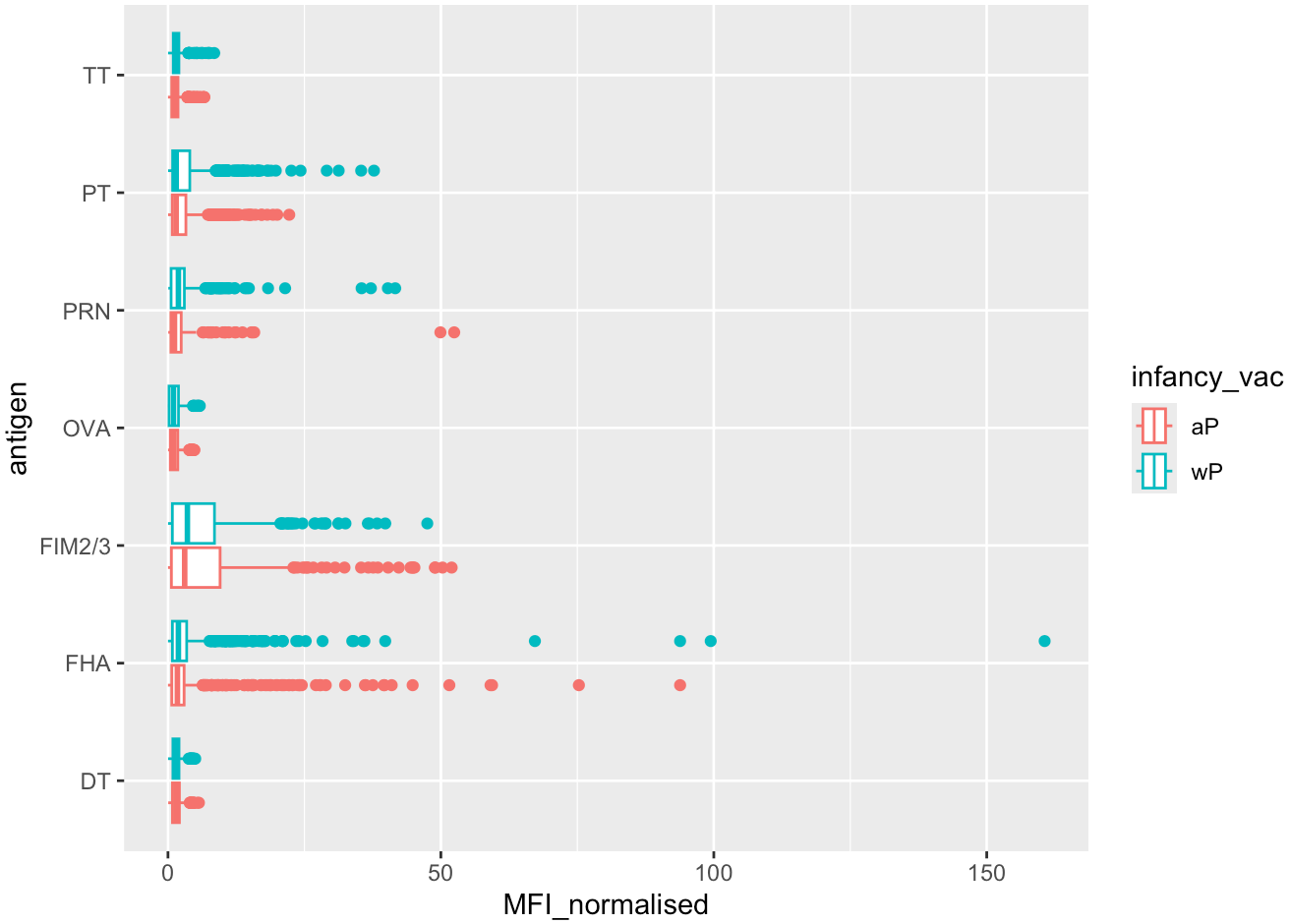


> Q16. What do you notice about these two antigens time courses and the PT data in particular?

The control antigen remains constant and low throughout all visits. The PT antigen increases drastically after the booster vaccines indicating a prevalent immune response.

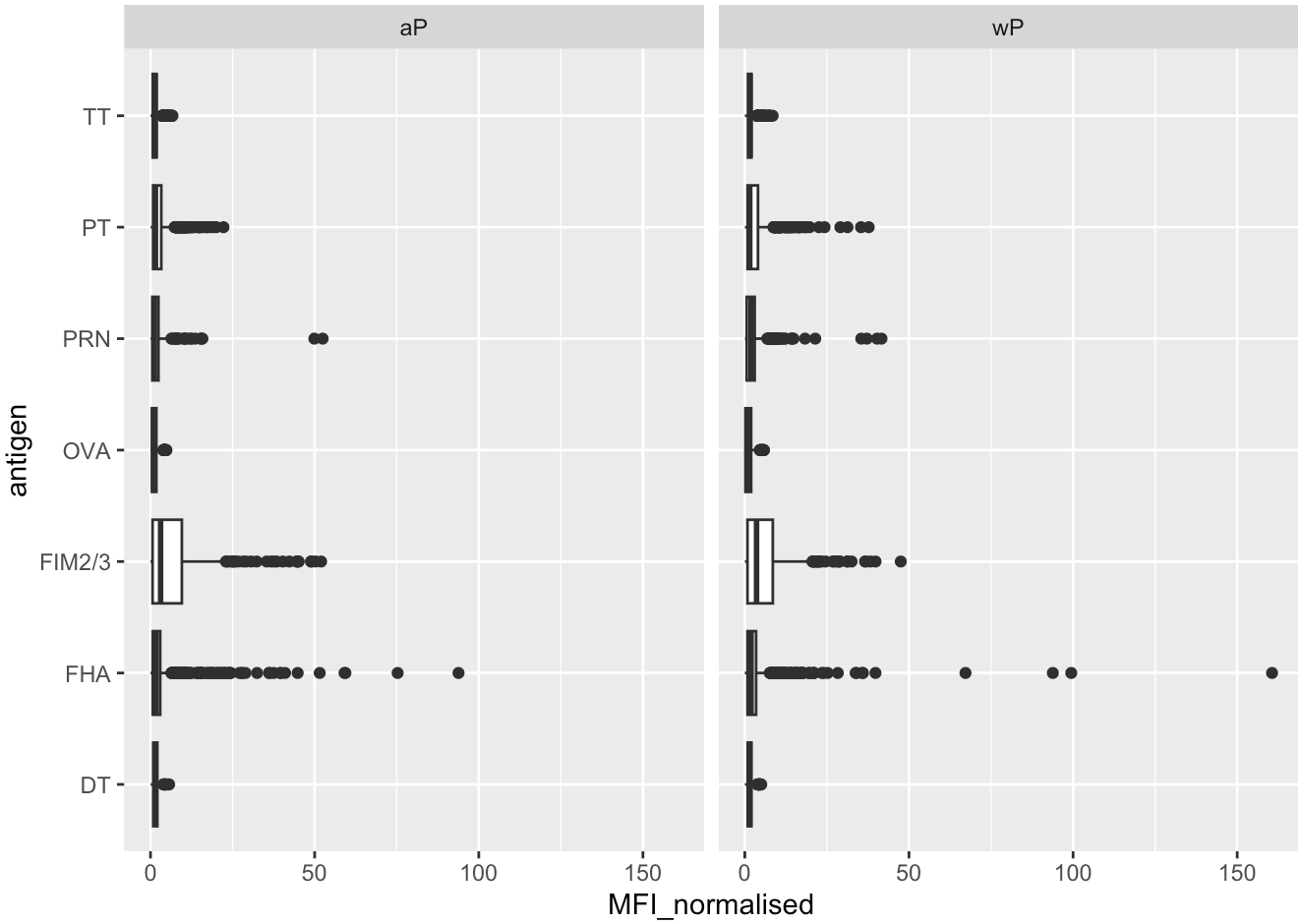
> Q17. Do you see any clear difference in aP vs. wP responses?

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



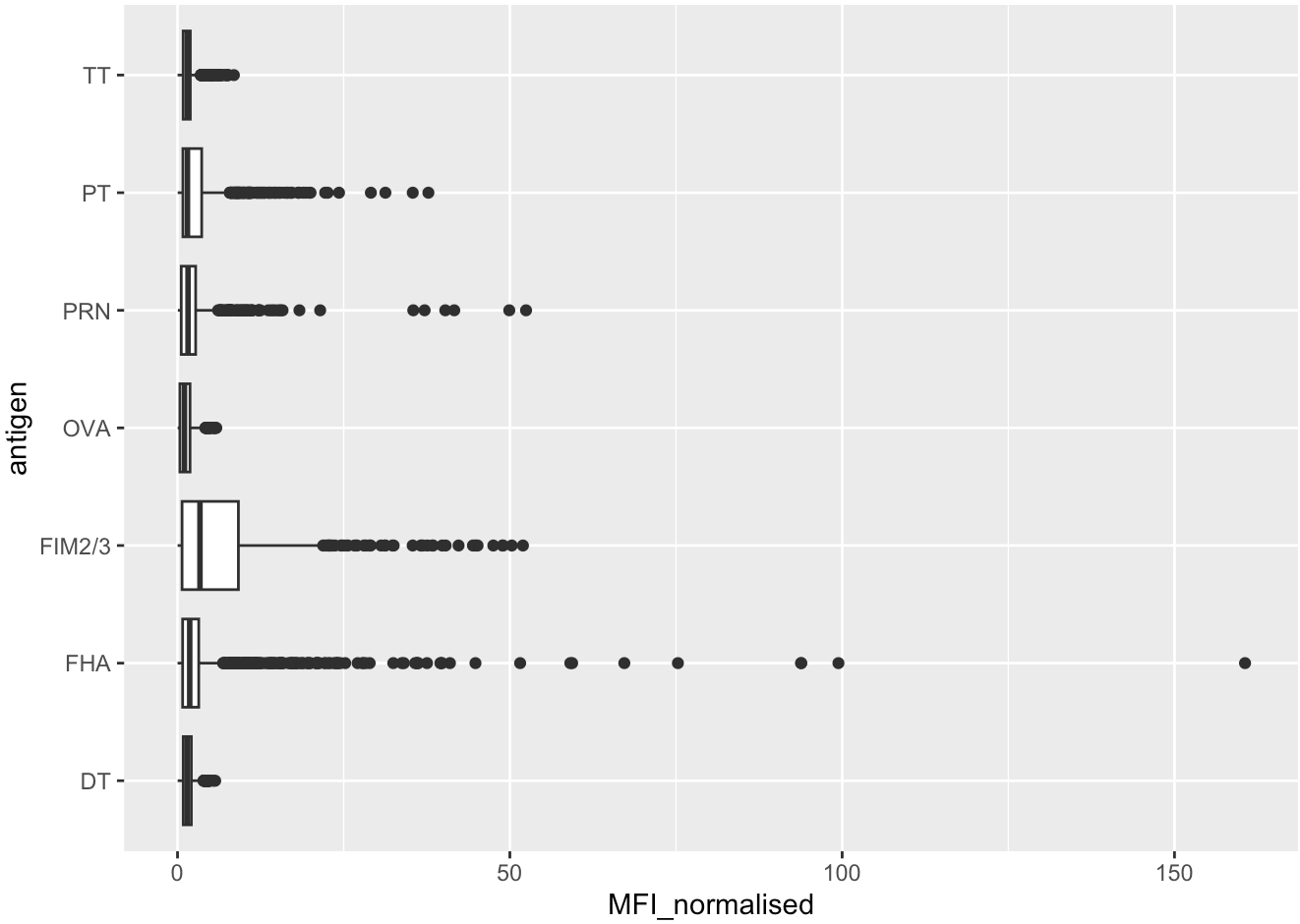
The wP group has higher PT antibodies than the aP group. Moreover, the wP group shows a stronger immune response to the vaccine booster than the aP group.

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



Is there a difference with time (ie.) before vs after booster shot?

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




```
    facet_wrap(~visit)
```

```
<ggproto object: Class FacetWrap, Facet, gg>
    attach_axes: function
    attach_strips: function
    compute_layout: function
    draw_back: function
    draw_front: function
    draw_labels: function
    draw_panel_content: function
    draw_panels: function
    finish_data: function
    format_strip_labels: function
    init_gtable: function
    init_scales: function
    map_data: function
    params: list
    set_panel_size: function
    setup_data: function
    setup_panel_params: function
    setup_params: function
    shrink: TRUE
    train_scales: function
```

```
    vars: function
    super:  <ggproto object: Class FacetWrap, Facet, gg>
```

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

```
abdata.21 <- abdata %>% 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)"
```