A Data visualisation project : Exploratory Data Analysis for SARS-COV-2 seropositivity

Audrey T Yeo, M Sci Biostats, M Nursing, RN, audreytyeo@gmail.com

Document was recreated in a reproducible format

Loading relevant packages

```r
library(readxl)
library(ggplot2)
library(janitor)
library(kableExtra)
library(tidyverse)
df = read.csv("serocov-pop_data_public.txt", sep = ",")
```

Data engineering

```r
set.seed(521)
names(df) #names of variables, there are 11 variables
```

```
##  [1] "ind_id"           "new_household_id" "bus_sante"        "Sex"
##  [5] "age"              "week"             "IgG_Ratio"        "pos"
##  [9] "neg"              "ind"              "IFA"
```

```r
dim(df) # 1000 x 11
```

```
## [1] 1000   11
```

```r
# data engineering
df$pos[df$pos == 1] = "positive"
df$pos[df$pos == 0] = "negative"
df$Sex[df$Sex == "1"] = "male"
df$Sex[df$Sex == "0"] = "female"
df$age = as.numeric(df$age)
names(df)[8] <- paste("FinalResult")
df$bmi = as.vector(sample(19:51, 1000, replace = TRUE))
n = dim(df)[1]
a = .13*n
b = .15*n
c = .10*n
d = .20*n
e = .03*n
f = .04*n
g = .01*n
h = .34*n
```

```r
df$comorbidities = c(rep("acute respiratory", a),
                     rep("chronic respiratory", b), rep("cancer", c),
                     rep("diabetes", d), rep("sedentary", e),
                     rep("other endocronology", f), rep("chronic renal", g),
                     rep("immunosupression", h))
```

Aim

The aim of the exploratory data analysis is to analyse the major risk factors for an infectious disease diagnosed bySARS-COV-2 seropositivy *based on synthetic data*

Descriptives on data

There are 1000 individuals in this study and 13 variables. There are 711 cases and 289 controls. Cases are defined as SARS-COV-2 seropositive denoted by outcome "positive" in variable "FinalResult". Across 13 variables, there are seven categories of comorbidities which can be understood as risk factors. Overall there are only 0 missing values for these categories. The cases represented approximately by the incidence of % of the study population. The breakdown of sexes are seen in the figure below. There is only one subject in a third sex group and this was a negative test. With respect to distribution in percentages, there is an equal proportion of male and female sexes within the cases and control group. A careful approach would be needed to impute and infer based on the available data for these variables.

Final result	male	female	total
negative	154	135	289
positive	366	345	711
Total	520	480	1000

Gender

Cases and controls by gender

The proportions of cases and controls by gender are as follows. Proportions are the same in each sex group for each final result.

Final result	count	percent
negative	0.29	0.29
positive	0.71	0.71

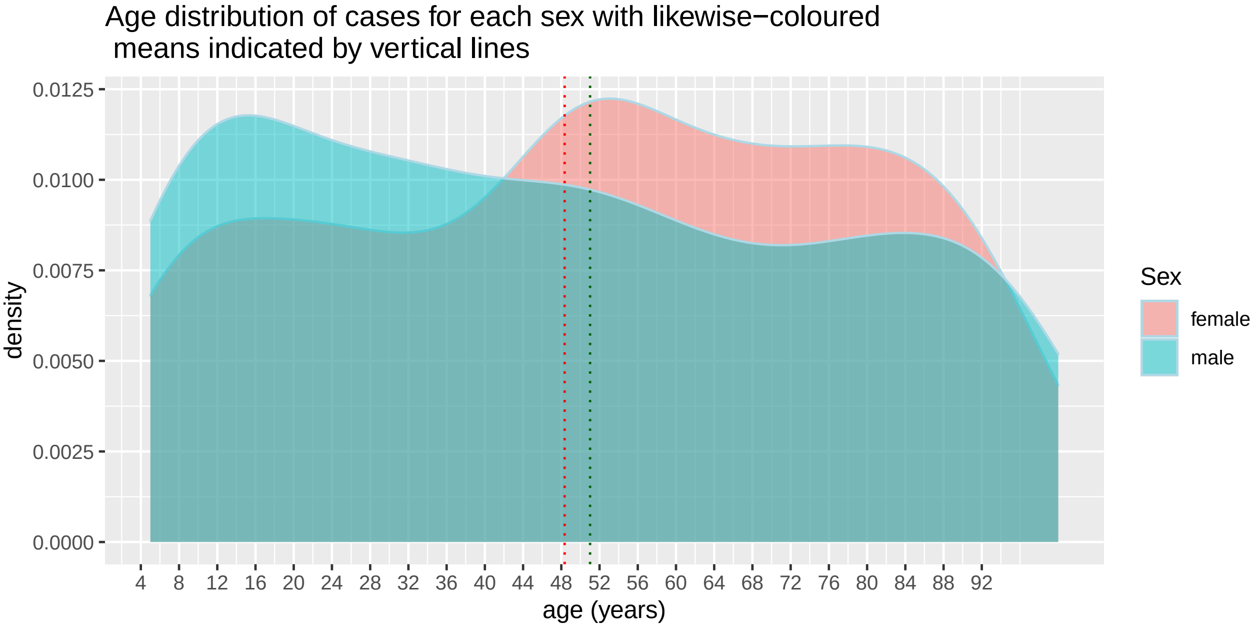
Age

Cases and controls by age and gender

Age across cases is (n = 711) and controls (n = 289). For the cases, males are on average 47.36 years old. The difference between male and female ages in cases are approximately -4.39 years younger than their female counterparts. Furthermore, the distribution of age across cases and controls are are 49.62 [5, 100] and controls 50.02 [5, 100].

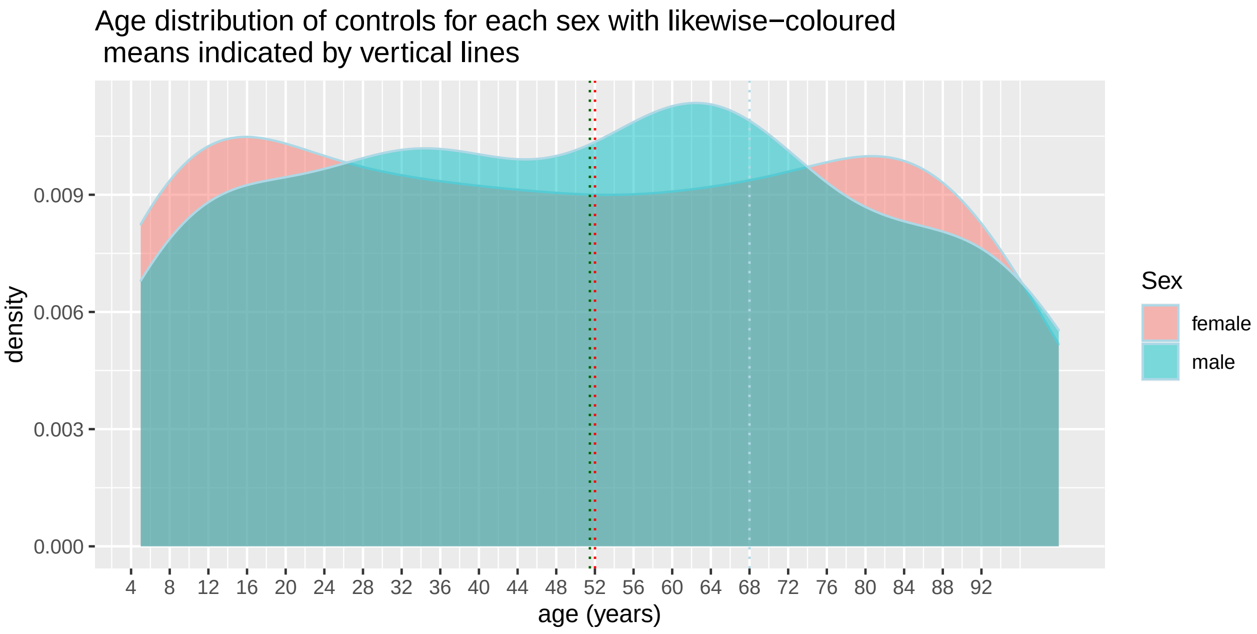
Cases by age and sex

The median age below and above 50 are NA for females and NA for males.



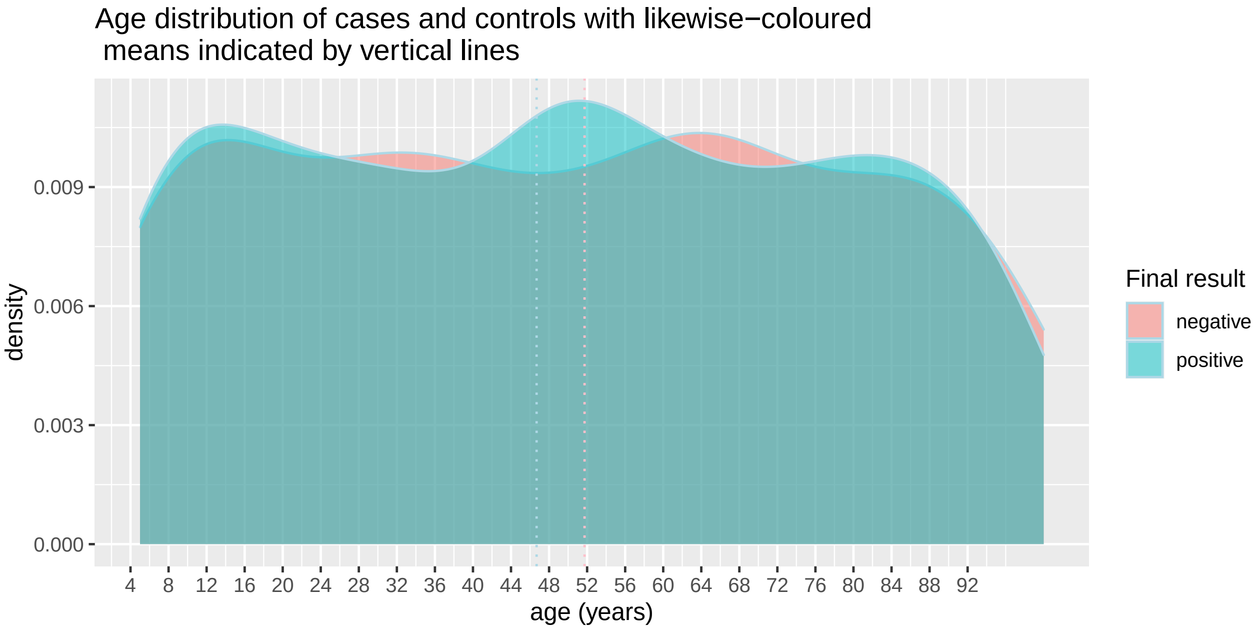
Controls by age and sex

The median age below and above 50 are 25 for females and 24 for males.


Age distribution of controls for each sex with likewise−coloured
means indicated by vertical lines

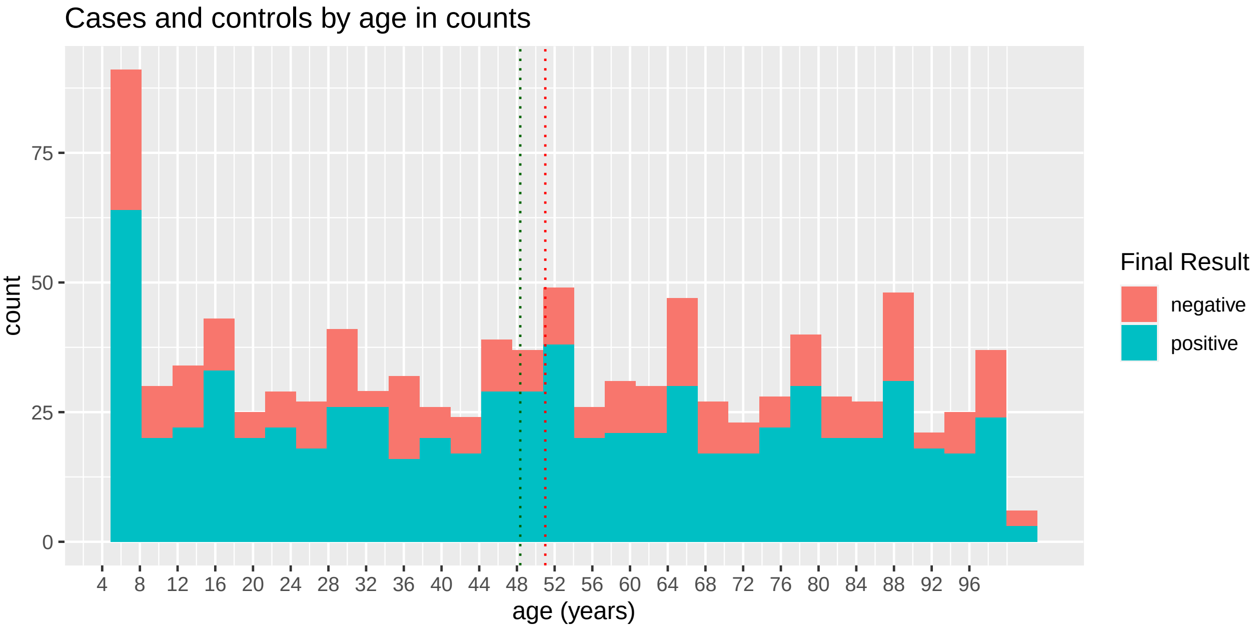
Cases and controls by age

Cases and controls are similarly distributed by age.

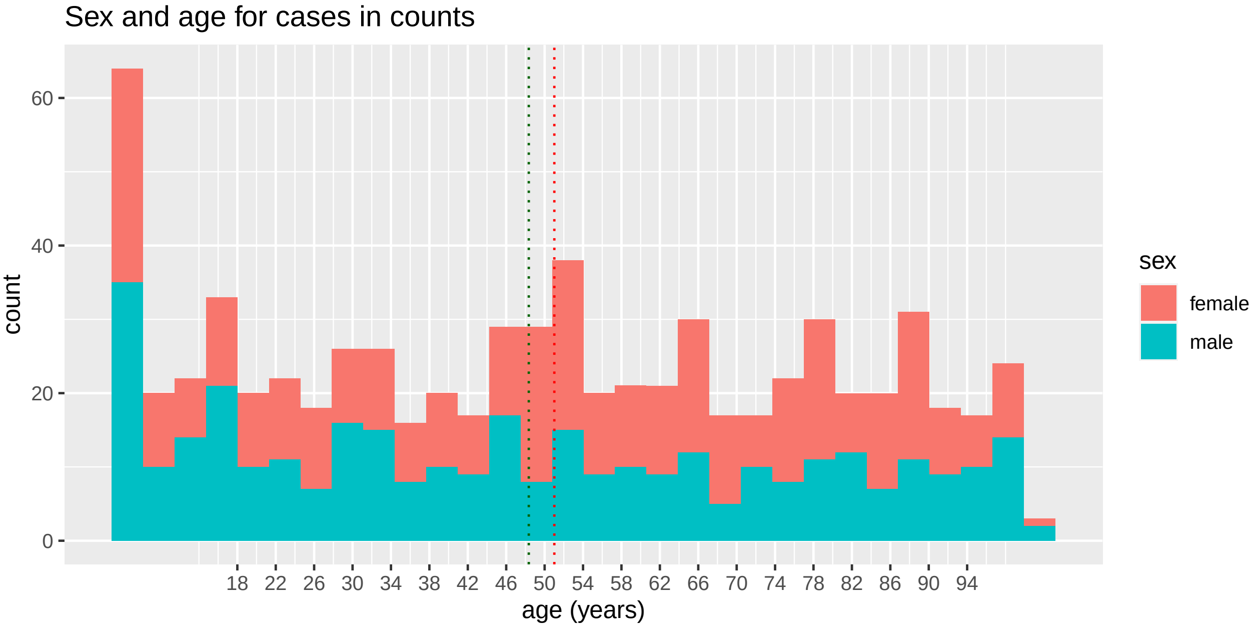

Age distribution of cases and controls with likewise−coloured
means indicated by vertical lines

Cases and controls by age in raw counts

Raw counts and distribution is seen below of age distribution in cases and controls.

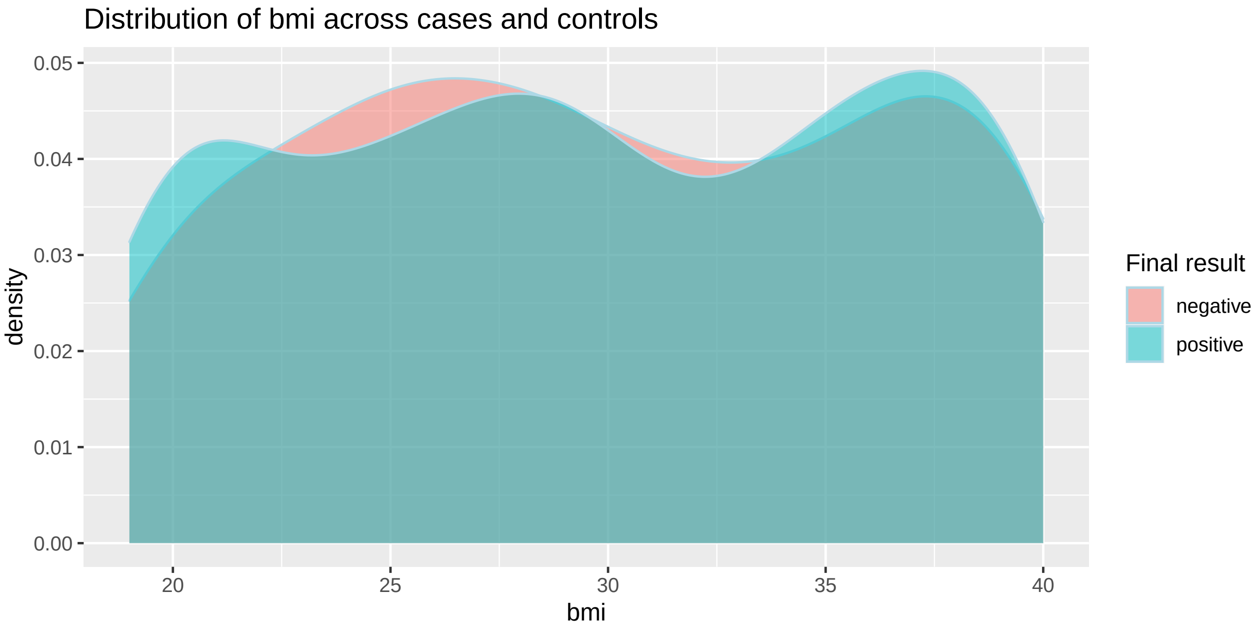


Cases by sex and age

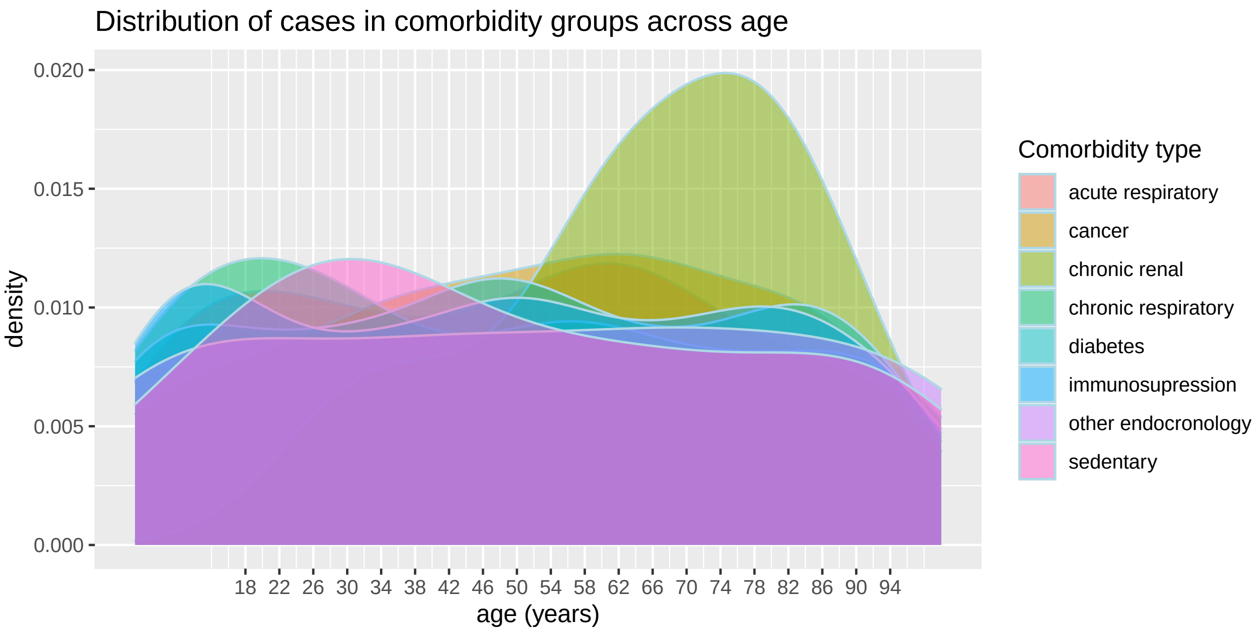


BMI

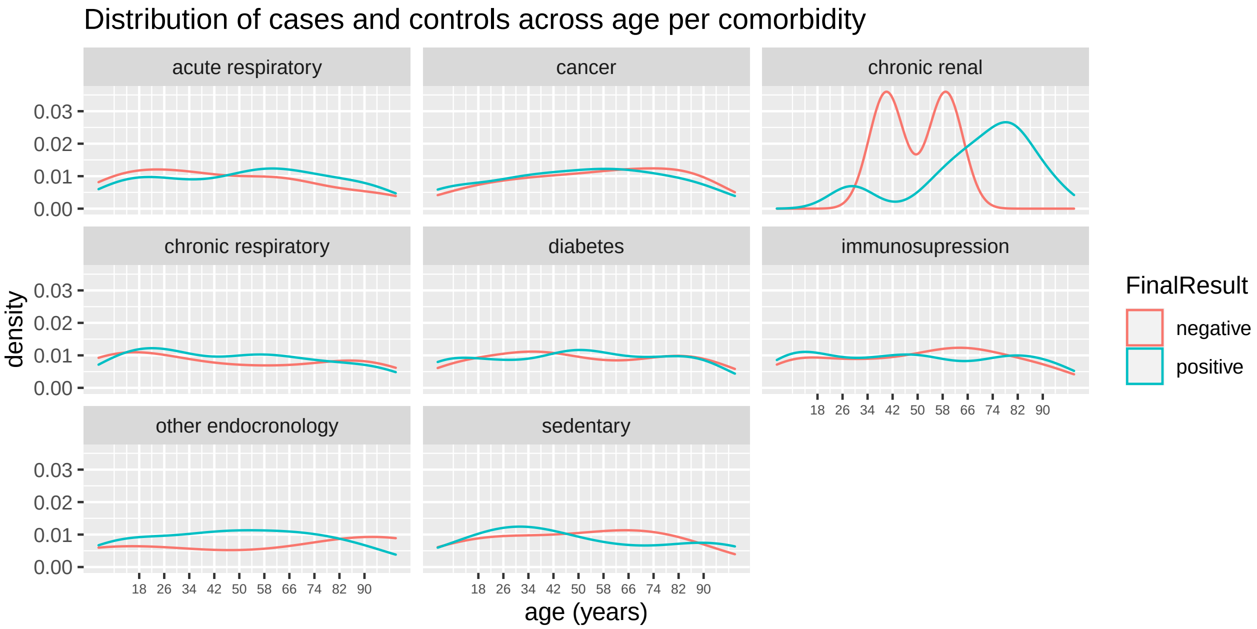
BMI across cases and controls



Comorbidity in cases

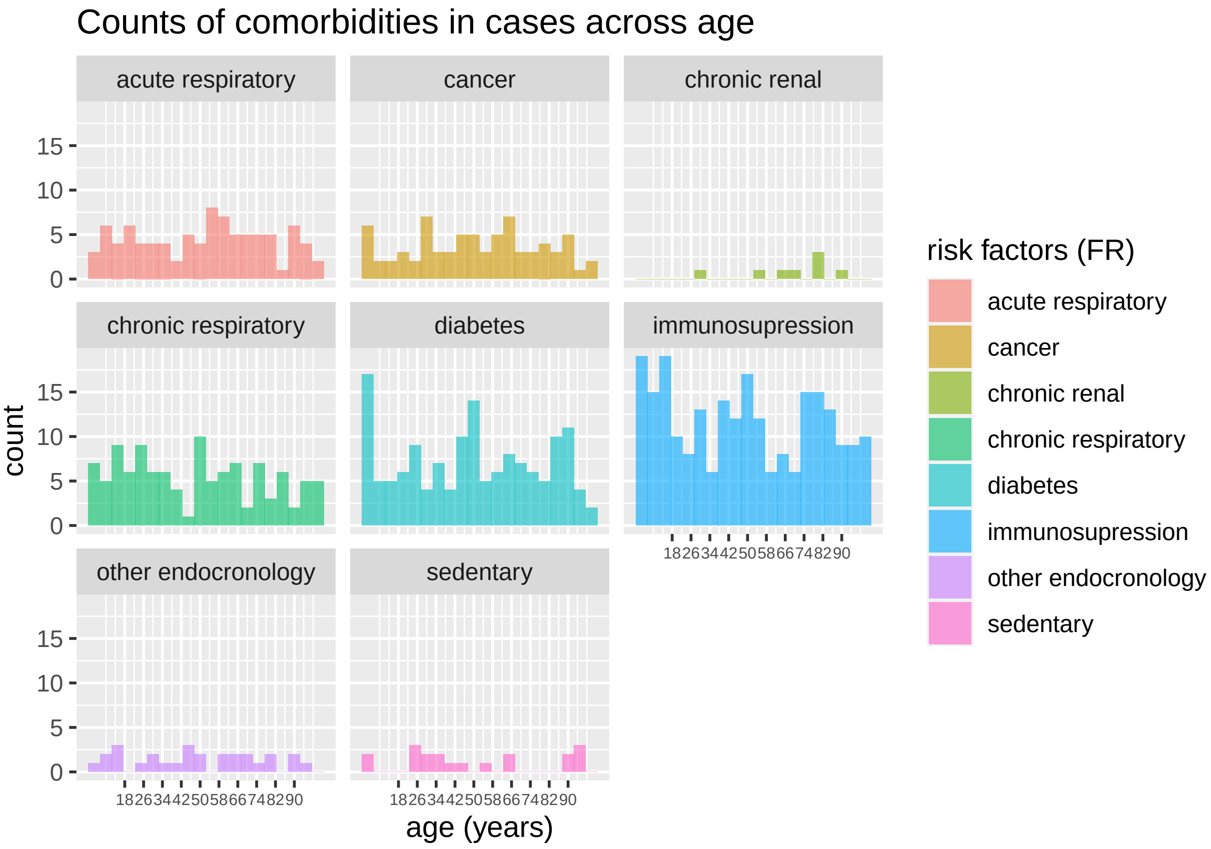


Cases and Controls distribution per comorbidity group



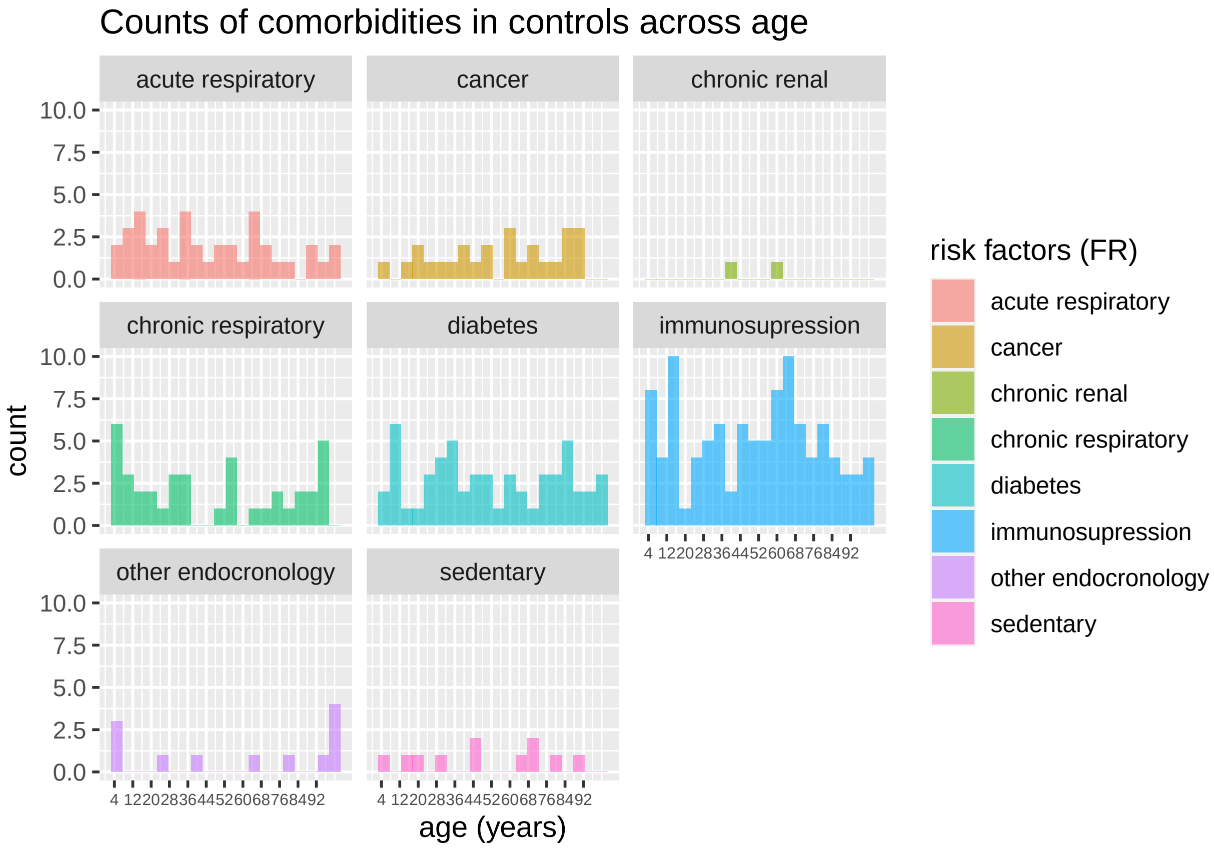
Comorbidities as risk factors in Cases

A raw account of comordities across age groups in cases is as follows.



Comorbidities as risk factors in Controls

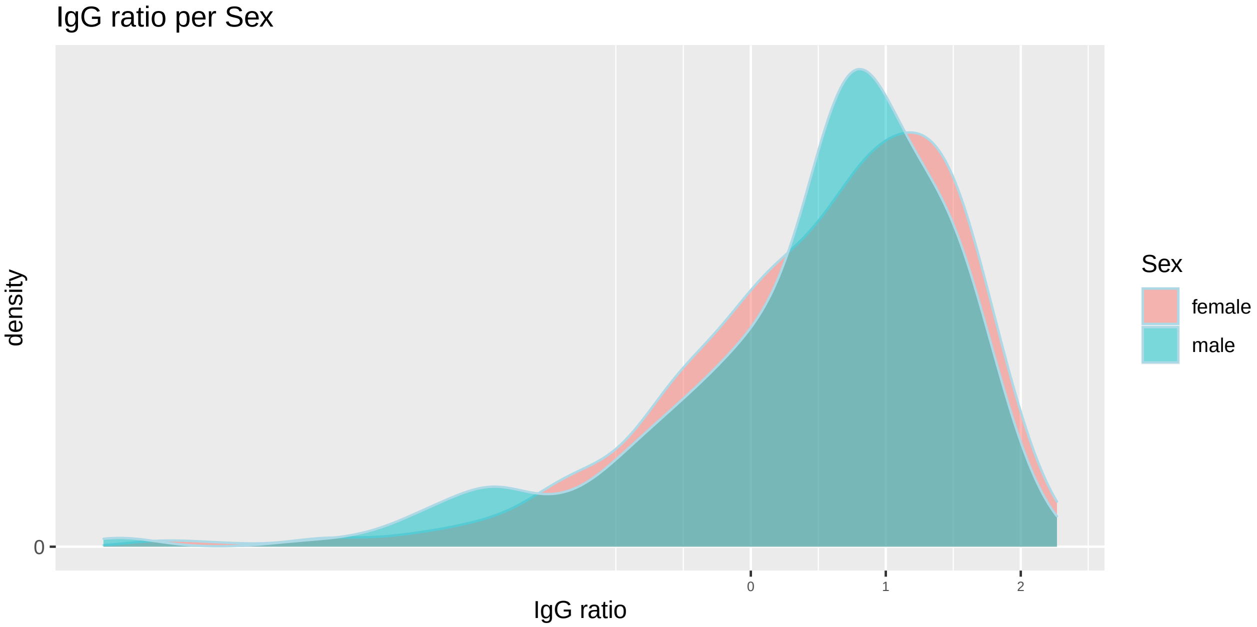
I compare across age groups in controls is as follows.



IgG

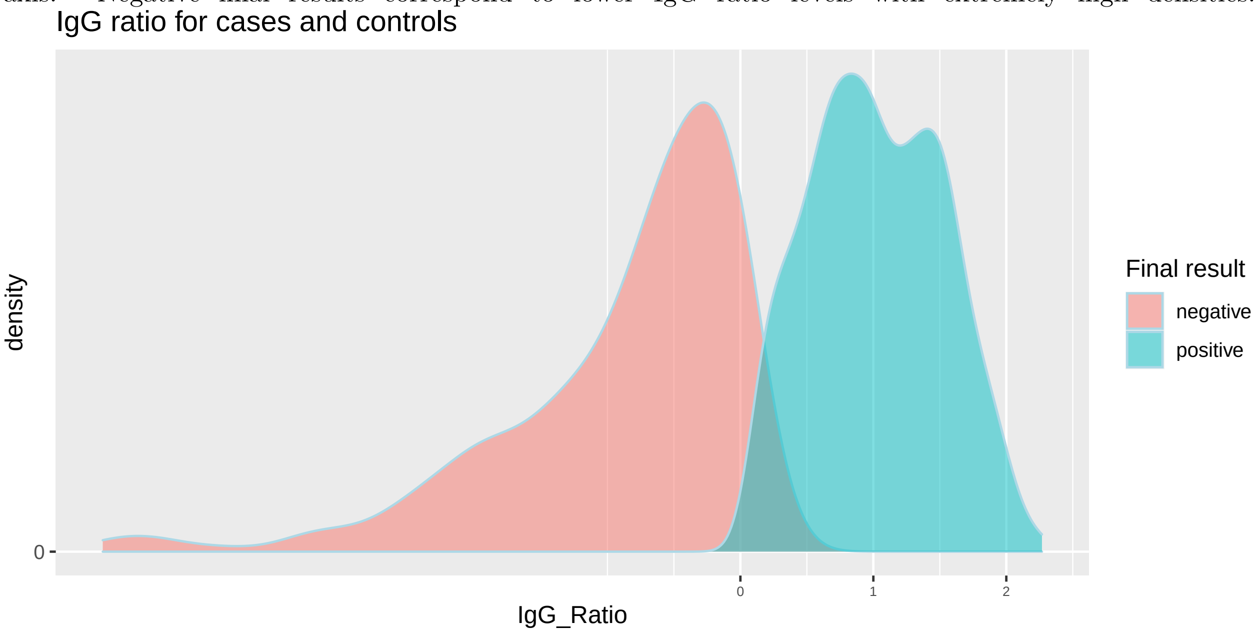
IgG distribution across IgG results

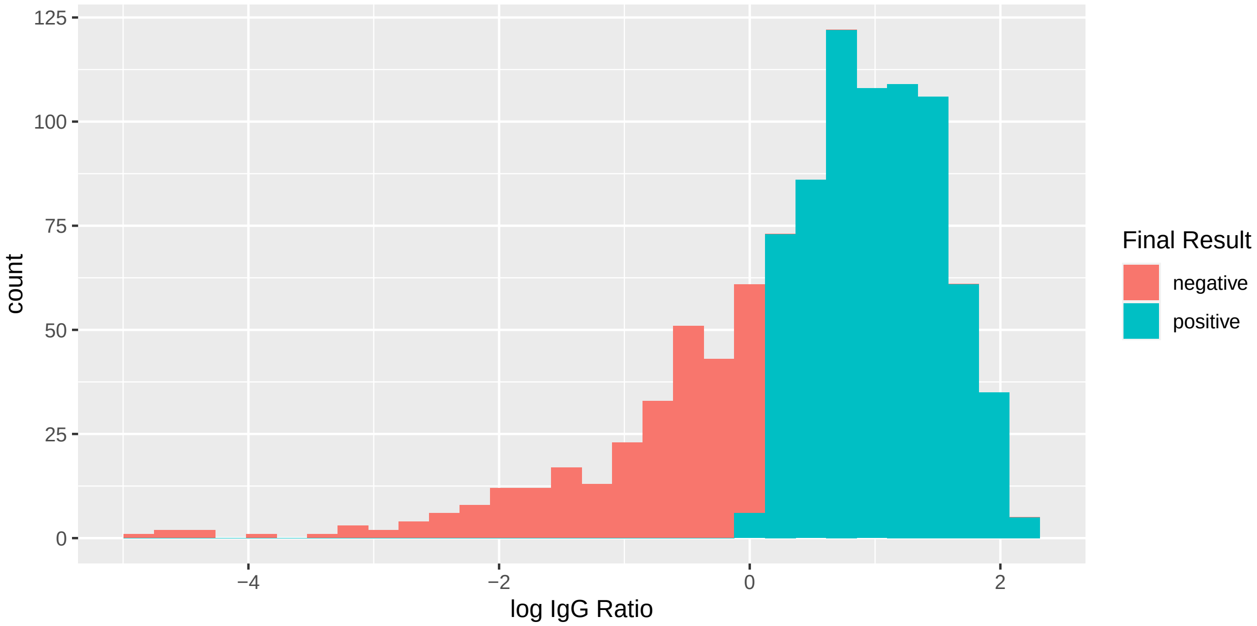
Important to note are IgG ratio levels and its correspondences to IgG results regardless of final results.



IgG for cases and controls

The below plot is the distribution of IgG ratio result of patients in cases and controls where there is a clearer distinction between a negative and positive final result. Positive final results correspond with higher IgG ratio levels although the peak is on the lower end on the x axis. Negative final results correspond to lower IgG ratio levels with extremely high densities.





Acknowledgements

Synthetic data was taken from https://github.com/HopkinsIDD/serocovpop/blob/master/generated_data/serocov-pop_data_public.csv

> I acknowledge the Traditional Aboriginal owners of land on which I find my privileges and endowments, blessings and responsibilities to my community, the Wurundjeri people of the Kulin Nations, and pay my respect to Elders past and present
>
> StopAsianHate