Exploratory Data Analysis for SARS-COV-2 seropositivity

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

```
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"
```

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

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

```
# 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$bmi = as.vector(sample(19:51, 1000, replace = TRUE))
n = dim(df)[1]
a = .13*n
b = .15*n
c = .10*n
d = .20*n
e = .03*n
f = .04*n
g = .01*n
h = .34*n
df$comorbidities = c(rep("acute respiratory", a), rep("chronic respiratory", b), rep("cancer", c), rep(
```

Aim

The aim of the exploratory data analysis is to analyse the major risk factors for an infectious disease diagnosed by SARS-COV-2 seropositivy.

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 852 missing values for these categories. The cases represented approximately by the incidence of 71% 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.

There are 13 variables surrounding the topic of travel location and dates, which has been posed as a risk factor to this infection. Overall 7% of data are missing from these variables. 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

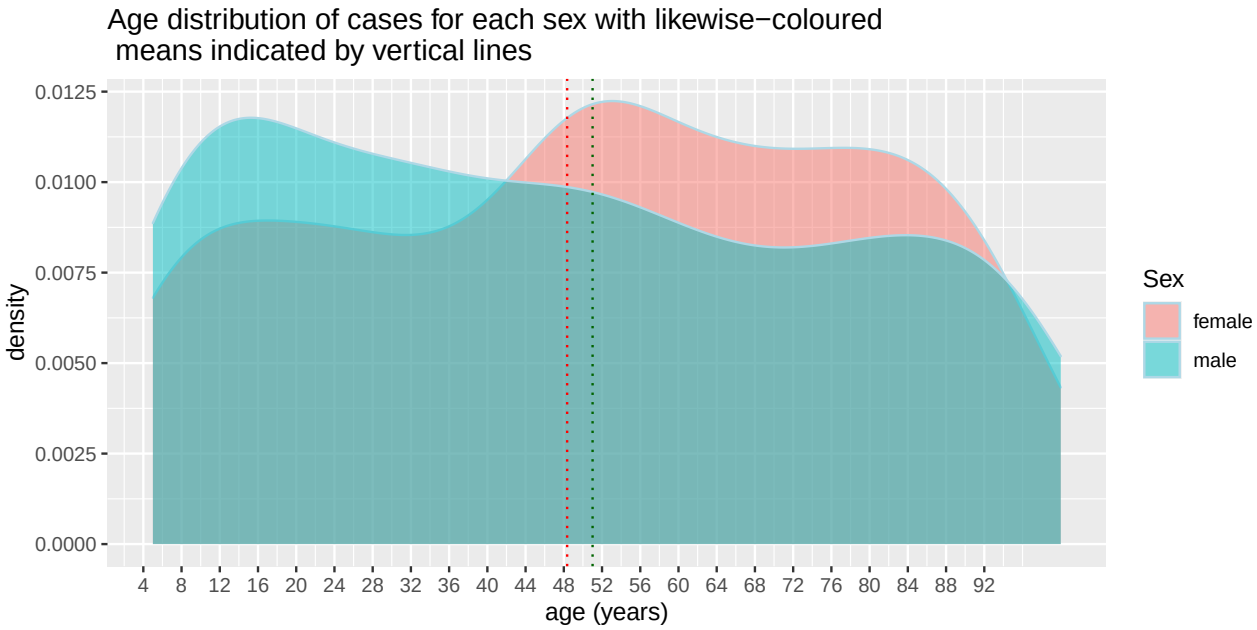
Age

Cases and controls by age and gender

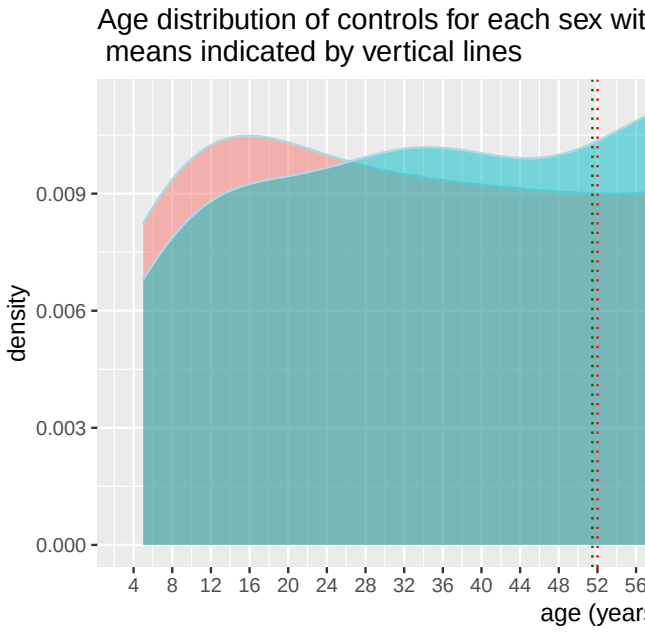
Age across cases is (n = 711) and controls (n = 289). For the cases, males are on average 47.36 years old. The difference between male and female ages in cases are approximately -4.39 years younger than their female counterparts. Furthermore, the distribution of age across cases and controls are are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



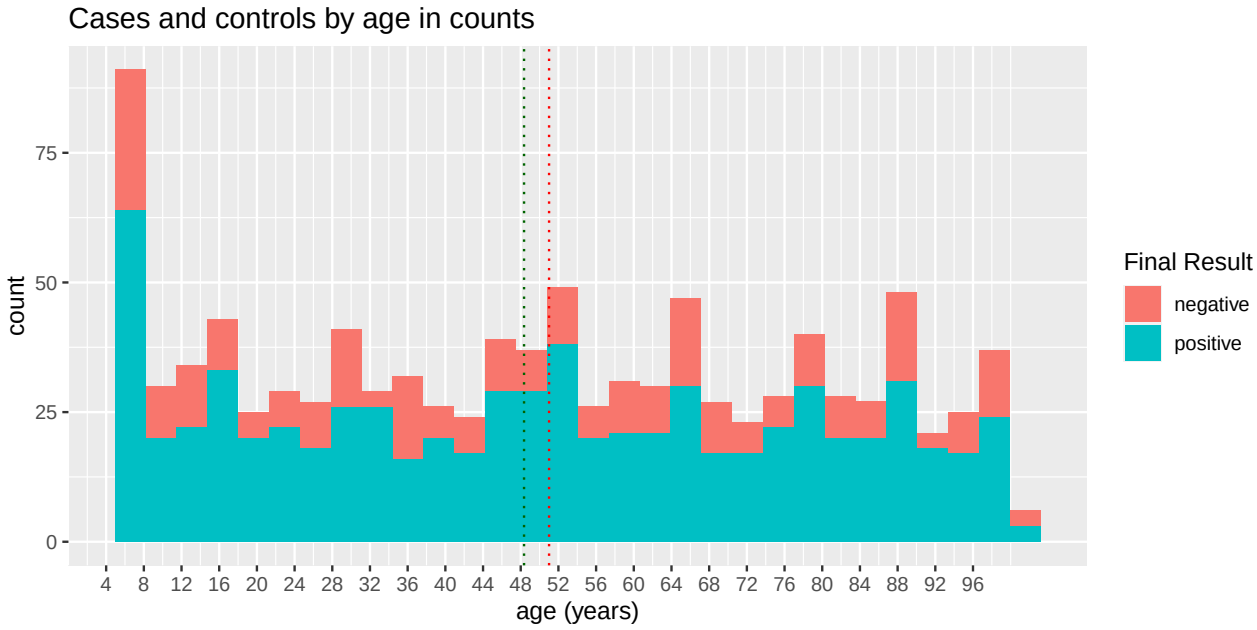
The median age below and above 50 are 25 for females and 24 for males.

Cases and controls by age

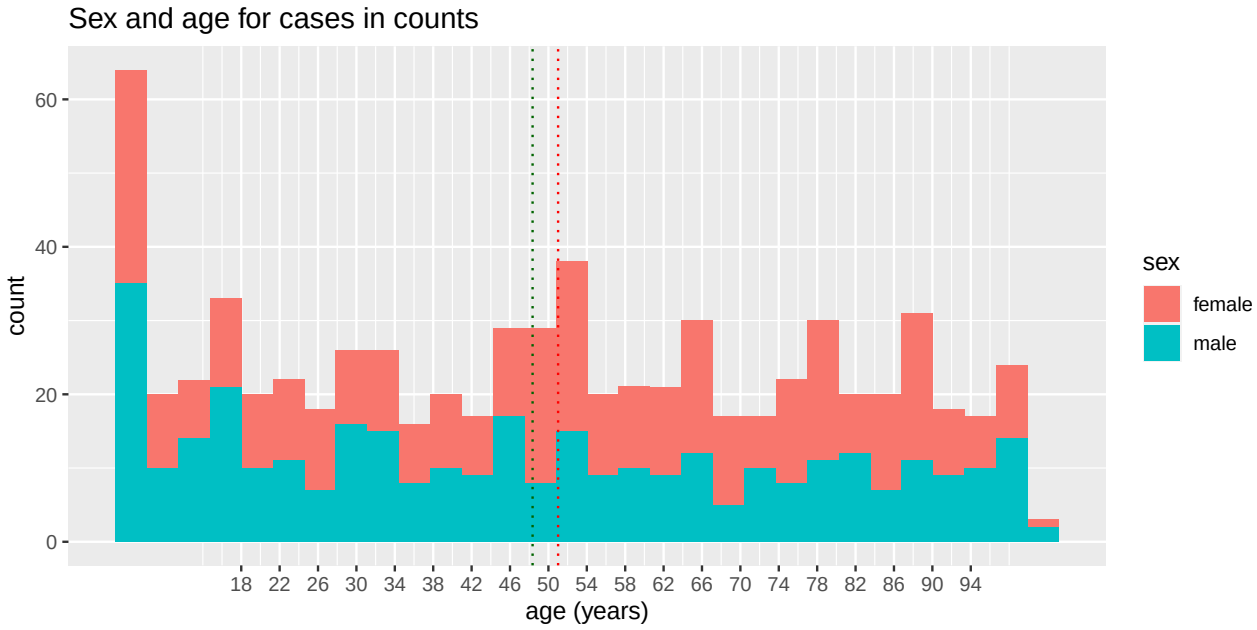
Cases and controls are similarly distributed by age. Two peaks occur at the age vicinity of 22 years and 50 years. Cases are heavily distributed in between ages of 38 and 62.

Cases and controls by age in raw counts

Raw counts and distribution is seen below of age distribution in cases and controls. The distribution pattern is similar as the peaks of the counts in the negative and positive final result groups. This could be a reflection of the nature of study participation which may favour certain age groups and or the nature of the spread and control of this infection.



Cases by sex and age



BMI

BMI data description

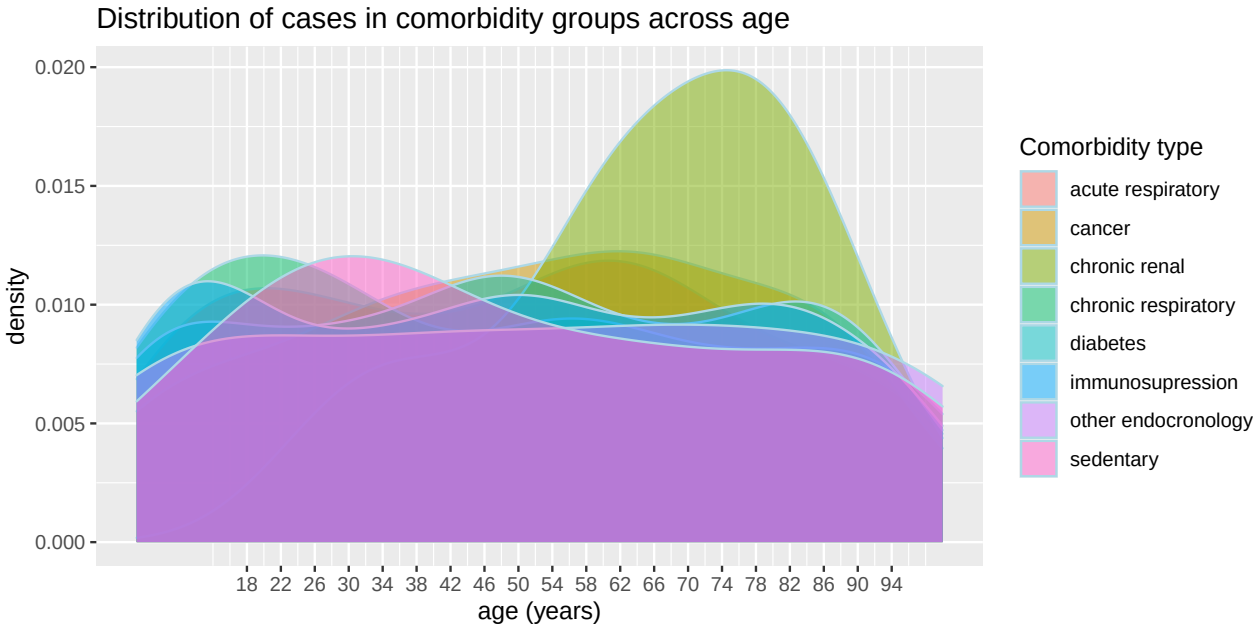
The BMI observations are between 21 to 340363 and it is understood that there are 289 observations above a BMI of 40. An example of extreme values are in the following table. These values do not correspond with clinical relevant values.

BMI across cases and controls

If however a clinically relevant cut off of BMI is 40, 681 observations are recorded, which reflects most of the observations. Below is the distribution in the cases and controls. The distribution is very similar across cases and controls.

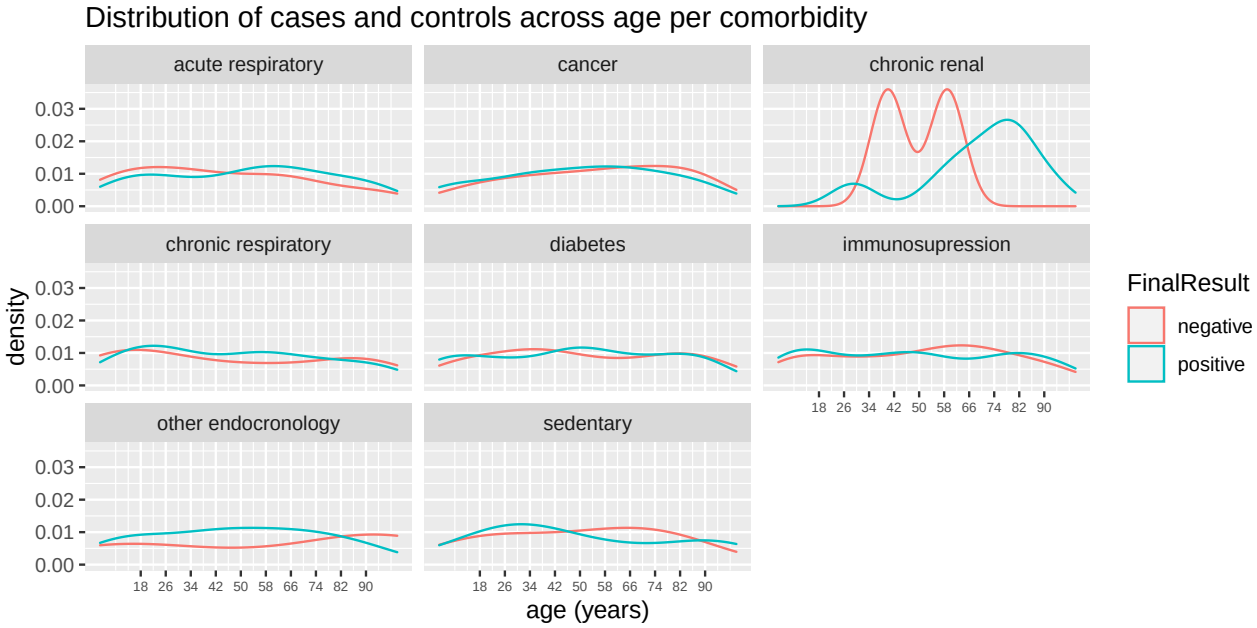
Comorbidity in cases

Most comorbidity groups peak in the ages of 50-78 years and follow similar behaviours at the x axis extremetires, while noting small densities in the case group.



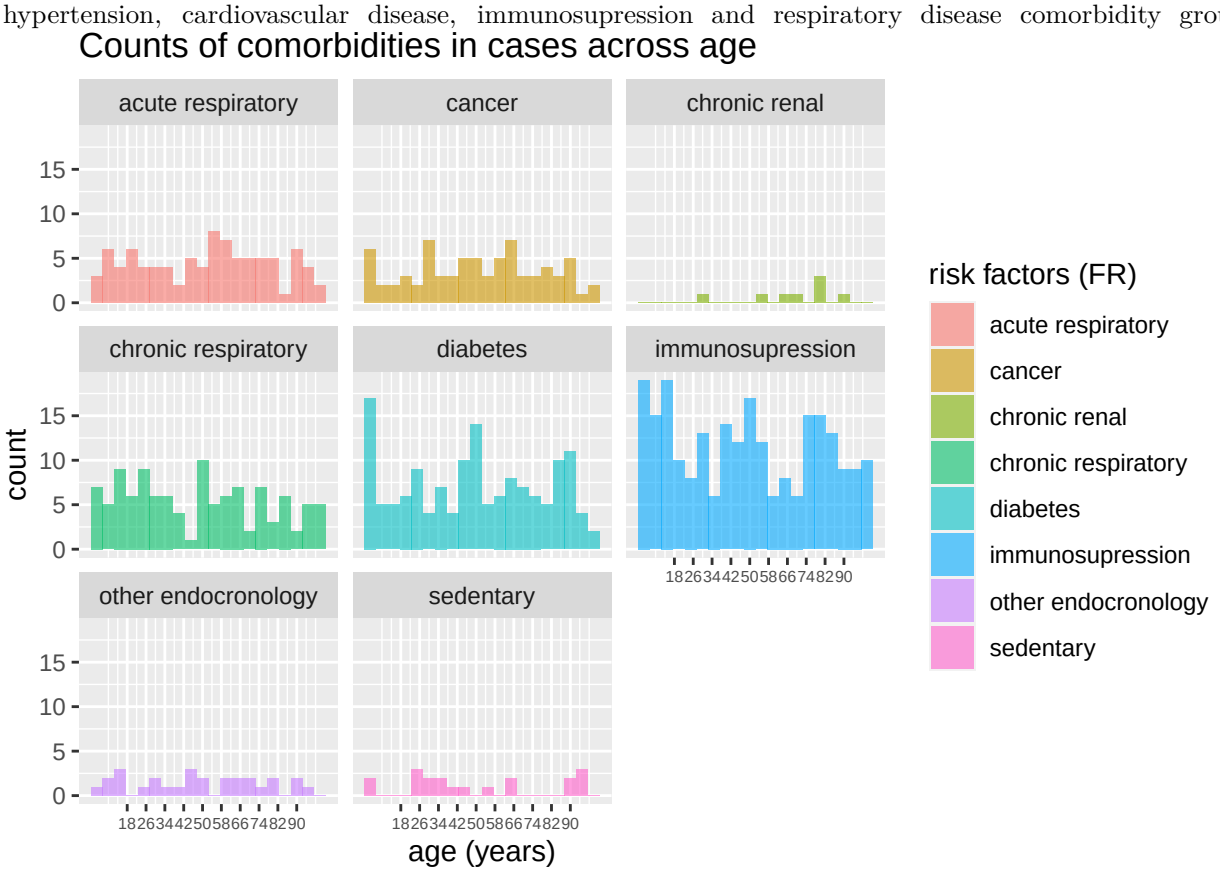
Cases and Controls distribution per comorbidity group

When compared, cases and controls have different distributions for each comorbidity group. Cases and controls have different peaks. The cases are more likely to have two peaks, for example in the hypertension, diabetes and respiratory disease groups which reflect two age vicinities where cases are highest in these respective comorbidities. Small densities of each are noted.



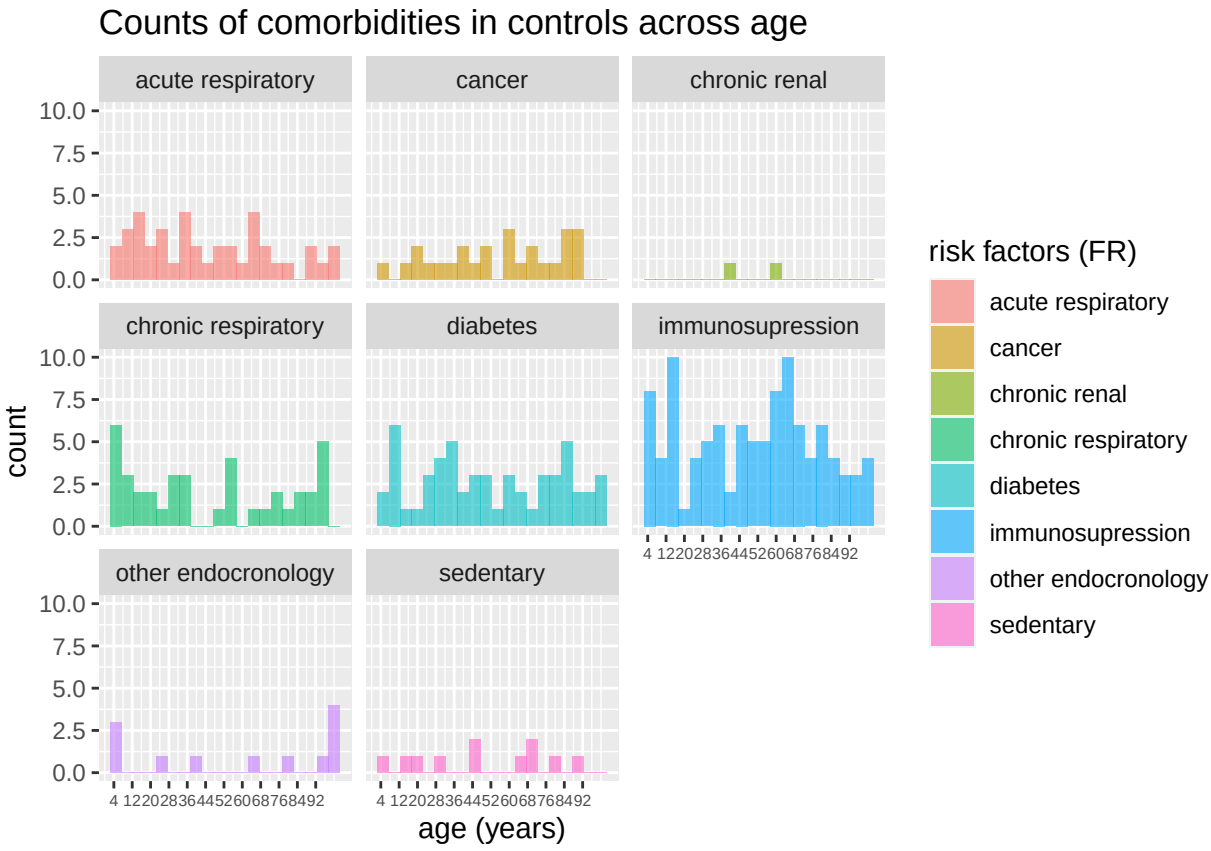
Comorbidities as risk factors in Cases

A raw account of comordities across age groups in cases is as follows. There are no similarities between groups even though an individual can be counted in more than one comorbidity group. There is a marked and consistent presence of comordbities in the age 50 and above group such as hypertension, cardiovascular disease, immunosupression and respiratory disease comorbidity group.



Comorbidities as risk factors in Controls

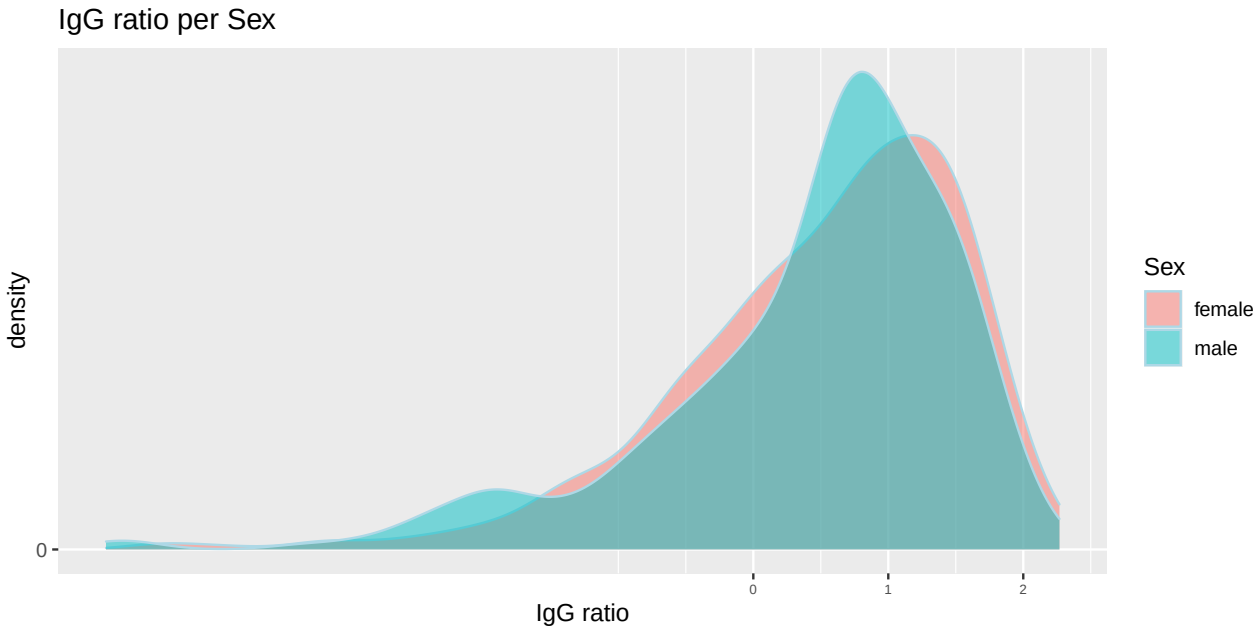
I compare across age groups in controls is as follows. There are similar trends between comorbidities as peaks are shared between the age groups of 50-70 years and a marked decrease on younger and older years on the x axis. This is comparably different from the cases comorbidity groups.



IgG

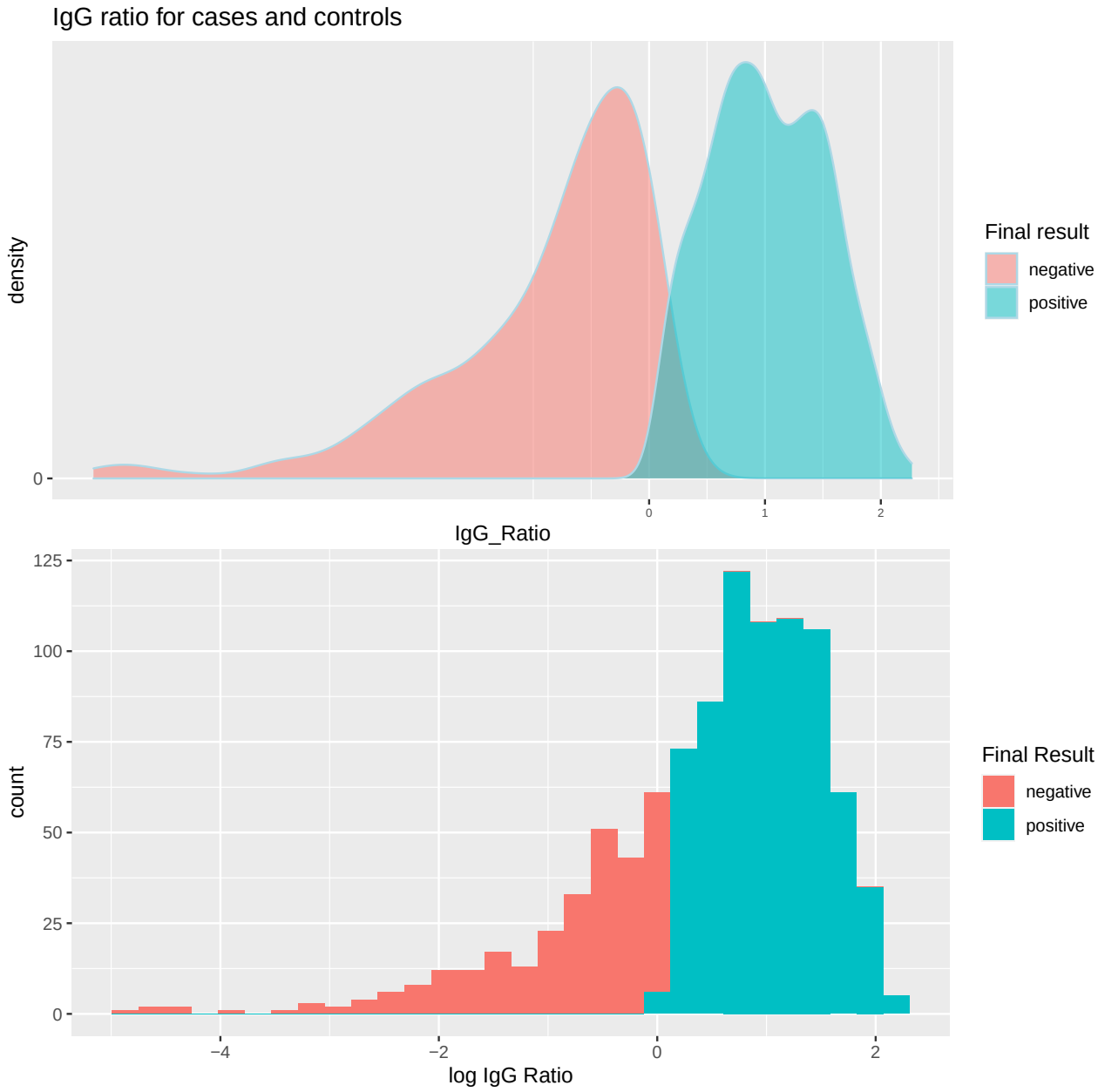
IgG distribution across IgG results

Important to note are IgG ratio levels and its correspondences to IgG results regardless of final results. Below is a graphical description of this association. IgG positive results usually correspond to higher IgG ratio levels and negative results to lower IgG ratio levels. There is a region of doubt or "dout" in between.



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.



Appendix

This document is in reproducible format.