Report on COVID19 Data

1/18/2022

Purpose

Obviously COVID19 has had huge impact globally right from the first case registered in late in 2019 in Wuhan, China until January 2022 at the time of writing. By now more than 5 million people have died from the disease and countless people have had their lives ruined. It is easy to think that this is just a continually downward spiral which just gets worse all the time. My impression was a different one however, as my impression was that there would be many subtrends and small victories that could be found when studying the numbers more closely.

Data used

The report will be based on the datasets "time_series_covid19_confirmed_global", "time_series_covid19_deaths_global", "time_series_covid19_recovered_global", "time_series_covid19_confirmed_US" and "time_series_covid19_deaths_US" by Johns Hopkins University. The datasets are all available on github at https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_series/.

```
# Import libraries first
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Vedhæfter pakke: 'lubridate'

## De følgende objekter er maskerede fra 'package:base':
##
##     date, intersect, setdiff, union
```

```
# Read data
base_url <-
  "https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_series/"
file_names <- c("time_series_covid19_confirmed_global.csv",
```

```
                "time_series_covid19_deaths_global.csv",
                "time_series_covid19_recovered_global.csv",
                "time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv")

urls <- str_c(base_url,file_names)

global_cases <- read_csv(urls[1])
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   `Province/State` = col_character(),
##   `Country/Region` = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
global_deaths <- read_csv(urls[2])
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   `Province/State` = col_character(),
##   `Country/Region` = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
global_recovered <- read_csv(urls[3])
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   `Province/State` = col_character(),
##   `Country/Region` = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
us_cases <- read_csv(urls[4])
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
```

```
##   Combined_Key = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
us_deaths <- read_csv(urls[5])
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Combined_Key = col_character()
## )
## i Use `spec()` for the full column specifications.
```

The raw data sets need to first be transformed a bit in order to do visualization and modelling on the data.

```
# Tidy data
global_cases <- global_cases %>%
  pivot_longer(cols = -c(`Province/State`,
                         `Country/Region`,
                         Lat,
                         Long),
               names_to = "date",
               values_to = "cases") %>%
  select(-c(Lat,Long))

global_deaths <- global_deaths %>%
  pivot_longer(cols = -c(`Province/State`,
                         `Country/Region`,
                         Lat,
                         Long),
               names_to = "date",
               values_to = "deaths") %>%
  select(-c(Lat,Long))

global_recovered <- global_recovered %>%
  pivot_longer(cols = -c(`Province/State`,
                         `Country/Region`,
                         Lat,
                         Long),
               names_to = "date",
               values_to = "recovered") %>%
  select(-c(Lat,Long))

us_cases <- us_cases %>%
  pivot_longer(cols = -c(UID,
                         iso2,
                         iso3,
```

```
                        code3,
                        FIPS,
                        Admin2,
                        Lat,
                        `Long_`,
                        Province_State,
                        Country_Region,
                        Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(-c(UID,
            iso2,
            iso3,
            code3,
            FIPS,
            Admin2,
            Lat,
            Long_,
            Province_State,
            Country_Region))

us_deaths <- us_deaths %>%
  pivot_longer(cols = -c(UID,
                        iso2,
                        iso3,
                        code3,
                        FIPS,
                        Admin2,
                        Lat,
                        `Long_`,
                        Province_State,
                        Country_Region,
                        Combined_Key,
                        Population),
               names_to = "date",
               values_to = "deaths") %>%
  select(-c(UID,
            iso2,
            iso3,
            code3,
            FIPS,
            Admin2,
            Lat,
            `Long_`,
            Province_State,
            Country_Region,
            Population))

global <- global_cases %>%
  full_join(global_deaths) %>%
  full_join(global_recovered) %>%
  rename(Country = `Country/Region`,
         Province = `Province/State`) %>%
```

```
  mutate(date = mdy(date))
```

```
## Joining, by = c("Province/State", "Country/Region", "date")
## Joining, by = c("Province/State", "Country/Region", "date")
```

```
us <- us_cases %>%
  full_join(us_deaths)  %>%
  mutate(date = mdy(date))
```

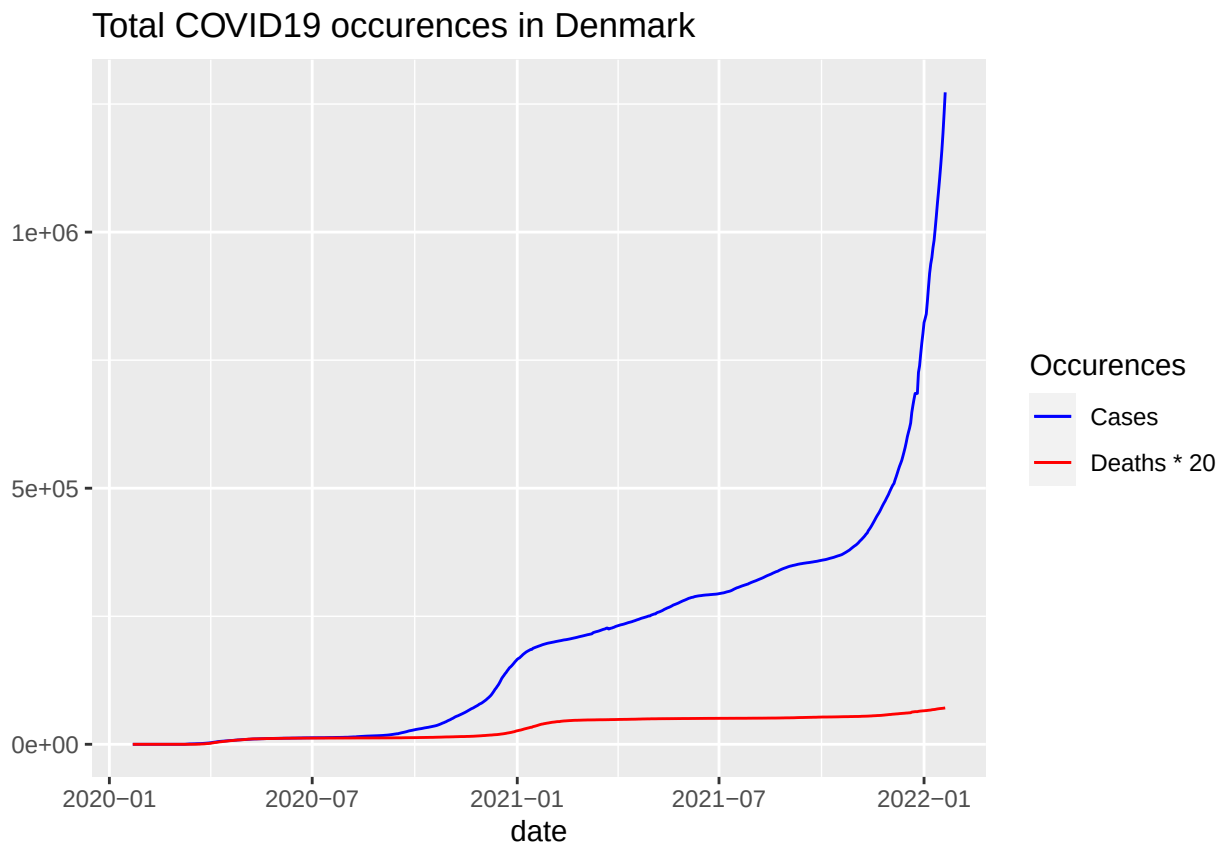
```
## Joining, by = c("Combined_Key", "date")
```

```
# Build dataset for aggregated counts for all chinese regions
china_totals <- global %>%
  filter(Country == "China") %>%
  group_by(date) %>%
  mutate(all_cases = sum(cases)) %>%
  mutate(all_deaths = sum(deaths)) %>%
  mutate(all_recovered = sum(recovered)) %>%
  filter(Province == "Hubei") %>%
  select(-c(Province,Country,cases,deaths,recovered))
```

COVID19 cases in Denmark

The diagram below shows the total number of COVID19 cases and deaths in Denmark (excluding Greenland and the Faroe Islands). The steep spike beginning late 2021 is due to the emergence of the Omicron variant (B.1.1.529) of the SARS-CoV-2 virus which has caused an exponential growth in the number of daily reported cases, however this large increase in cases due to Omicron does not seem to reflect in a significantly higher number of deaths (yet at least), indicating that the Omicron variant might be less lethal than the Delta variant, at least in combination with the danish vaccination effort. The effects of the vaccination efforts and general restrictions put in place to combat the virus can be seen already from the second half of 2020. It should be noted that deaths are shown at a different scale than cases.

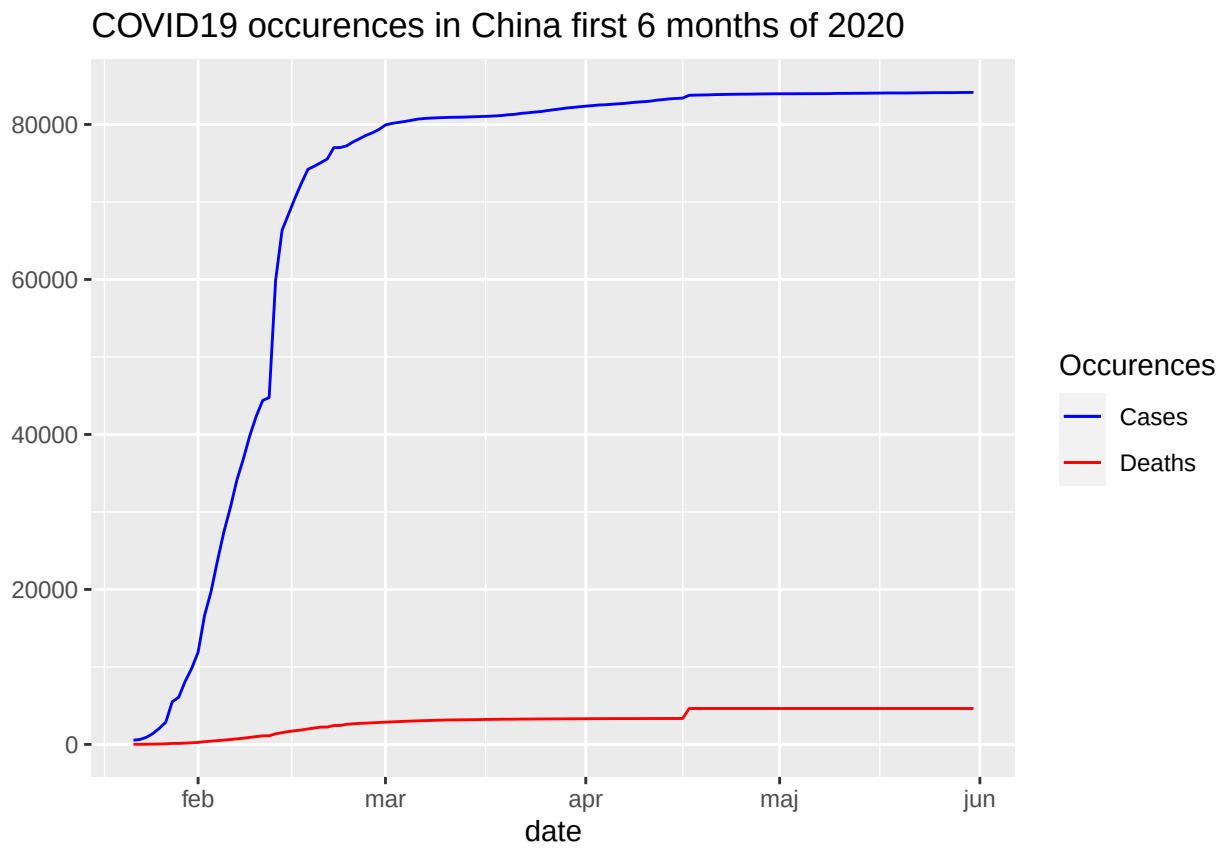
```
global %>%
  filter(Country == "Denmark", is.na(Province)) %>%
  ggplot(aes(x = date)) +
  geom_line(aes(y=cases, color="Cases")) +
  geom_line(aes(y=deaths * 20,color="Deaths * 20")) +
  scale_color_manual(name = "Occurences",
                     values = c("Cases" = "blue",
                                "Deaths * 20" = "red")) +
  labs(title = "Total COVID19 occurences in Denmark", y=NULL)
```



Chinese occurences early in the pandemic

It is generally agreed that the SARS-CoV-2 virus first appeared in the Chinese city of Wuhan mid- or late 2019. The dataset provided does not cover the first months after discovery of the virus but it is interesting to see that even though the Chinese efforts to combat the virus were criticized for being heavy-handed they also proved to be very effective as the number of new cases become very low by the beginning of march 2020.

```
china_totals %>%
  filter(date < as.Date("2020-06-01")) %>%
  ggplot(aes(x = date)) +
  geom_line(aes(y= all_cases, color = "Cases")) +
  geom_line(aes(y= all_deaths, color = "Deaths")) +
  scale_color_manual(name = "Occurences",
                     values = c("Cases" = "blue",
                                "Deaths" = "red")) +
  labs(title = "COVID19 occurences in China first 6 months of 2020", y=NULL)
```



COVID19 trend

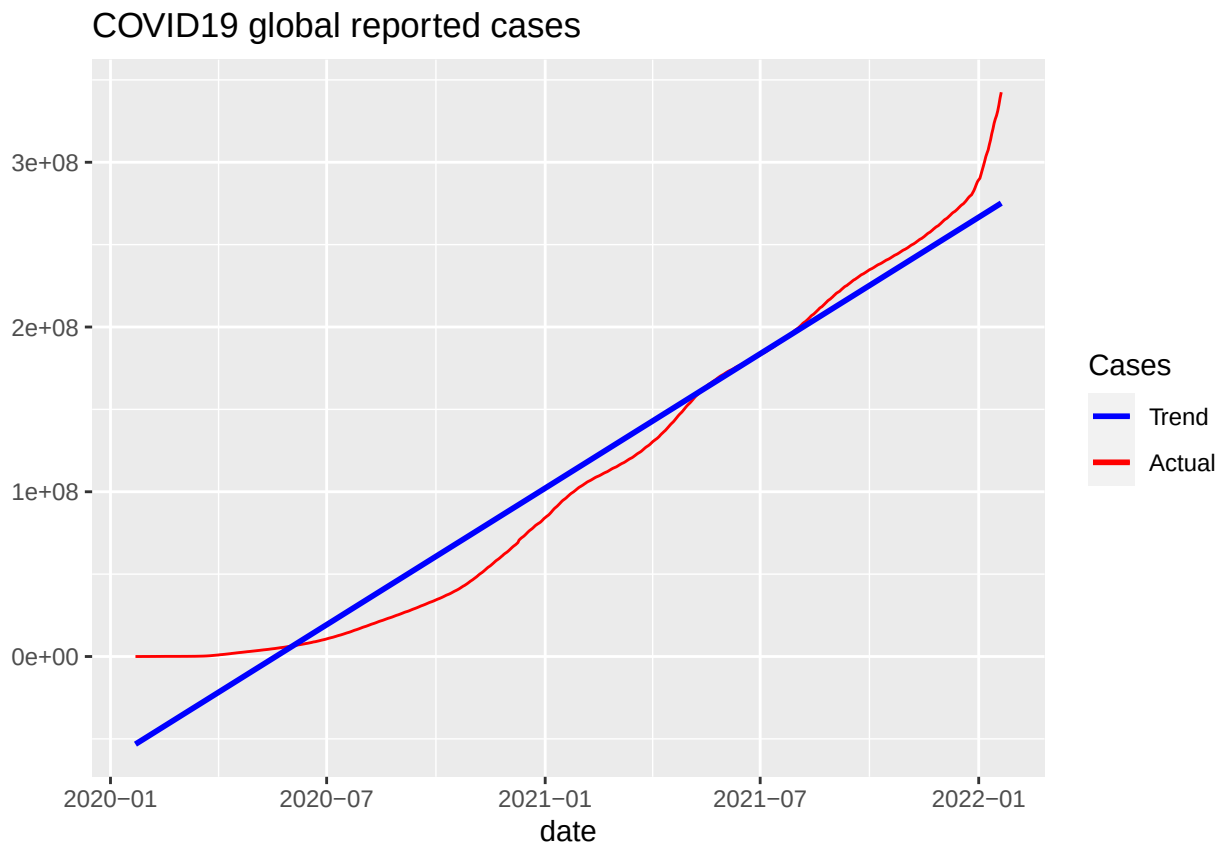
The first of the two diagrams below shows the global development in COVID19 case with a linear model fitted to show the trend. COVID19 cases were rising slowly the first half of 2020 as the virus was still relatively localized and then significantly increased after it had more or less spread all over the world and thus become a global pandemic. From around new year 2022 we see an even steeper incline in global COVID19 cases, likely due to the emergence of more and more contagious variants of the virus.

The second diagram below shows the relation between cases and deaths (how deadly the virus is at a given point of time from a global perspective). It is very clear that while the virus was not very wide-spread during early 2020 it was significantly more deadly than later. From around mid-2020 until around early 2021 there is a steep decline in deadliness of the virus, coinciding with the appearance of different vaccines targeted towards SARS-CoV-2. An additional pattern worth noticing is that from very late 2021 there is a spike in occurences seen in the first diagram and a drop in deadliness as seen on the second diagram. Scientists have speculated that this is due to the emergence of the Omicron variant of the virus (B.1.1.529) which is hoped to be less lethal but more effective in spreading between vaccinated or previously infected people.

```
global_sum <- global %>%
  drop_na(cases) %>%
  group_by(date) %>%
  mutate(all_cases = sum(cases)) %>%
  mutate(all_deaths = sum(deaths)) %>%
  select(-c(Province,Country,cases,deaths,recovered))

global_sum %>%
  ggplot(aes(x = date, y = all_cases, color="Trend")) +
           geom_line(aes(color="Actual")) +
```

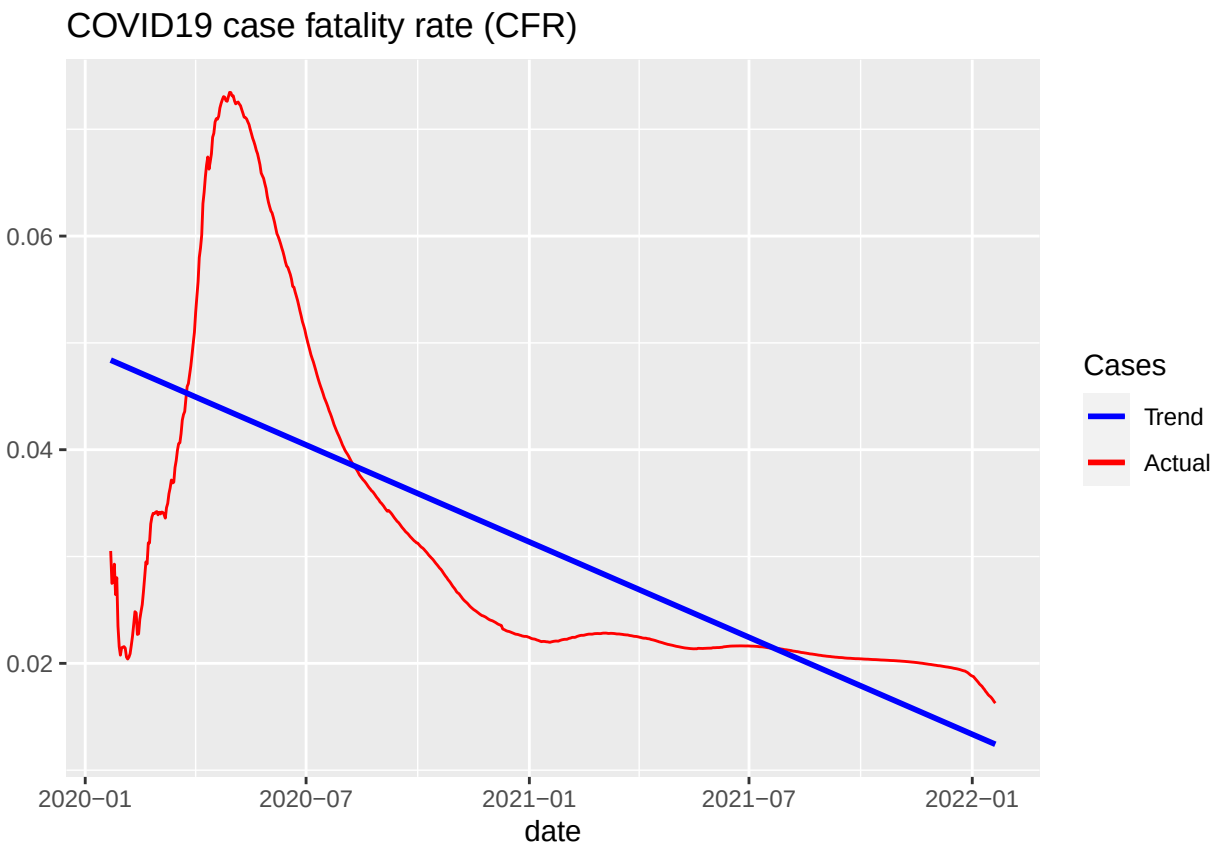
```
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_manual(name = "Cases",
                     values = c("Trend" = "blue",
                                "Actual" = "red")) +
  labs(title = "COVID19 global reported cases", y=NULL)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
model <- lm(formula = all_deaths ~ all_cases, data = global_sum)

global_sum %>%
  ggplot(aes(x = date, y = all_deaths / all_cases, color="Trend")) +
           geom_line(aes(color="Actual")) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_manual(name = "Cases",
                     values = c("Trend" = "blue",
                                "Actual" = "red")) +
  labs(title = "COVID19 case fatality rate (CFR)", y=NULL)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Fatality model

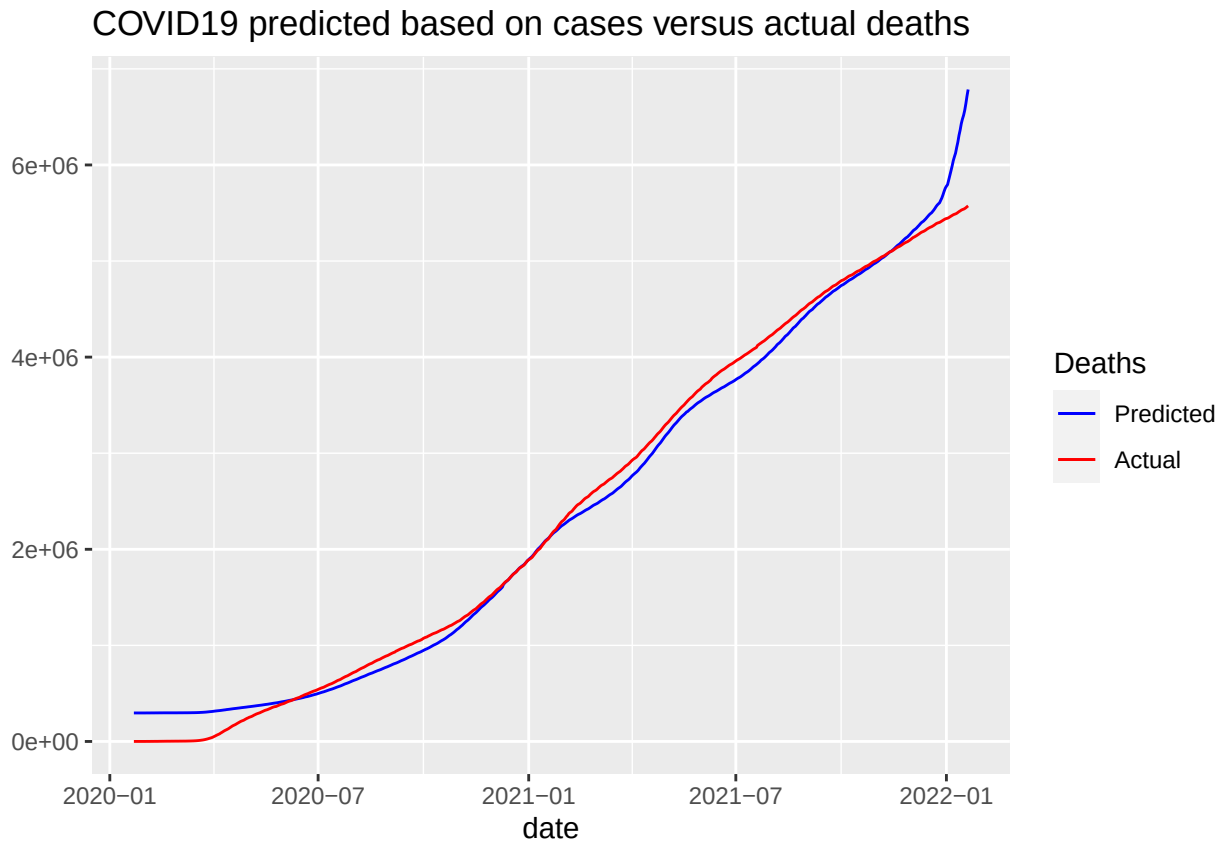
With the diagram below I would like to show is a linear model fitted entirely to deaths as a function of cases. As can be seen in the diagram the total deaths predicted fit rather closely with the actual death counts. Again we see the steep spike in 2022 where the predicted deaths are significantly higher than the actual deaths, indicating that the Omicron variant is less lethal than especially the Delta variant which is responsible for many deaths in 2021.

```
model <- lm(formula = all_deaths ~ all_cases, data = global_sum)
summary(model)
```

```
##
## Call:
## lm(formula = all_deaths ~ all_cases, data = global_sum)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1211436   -36221    51100   121107   194381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.965e+05  6.393e+02   463.8   <2e-16 ***
## all_cases   1.895e-02  4.338e-06  4367.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 190100 on 204398 degrees of freedom
## Multiple R-squared:  0.9894, Adjusted R-squared:  0.9894
## F-statistic: 1.908e+07 on 1 and 204398 DF,  p-value: < 2.2e-16
```

```
global_sum %>%
   ggplot(aes(x = date)) +
   geom_line(aes(y=predict(model),color="Predicted")) +
   geom_line(aes(y=all_deaths,color="Actual")) +
   scale_color_manual(name = "Deaths",
                      values = c("Predicted" = "blue",
                                 "Actual" = "red")) +
  labs(title = "COVID19 predicted based on cases versus actual deaths", y=NULL)
```



US Vaccination effectiveness

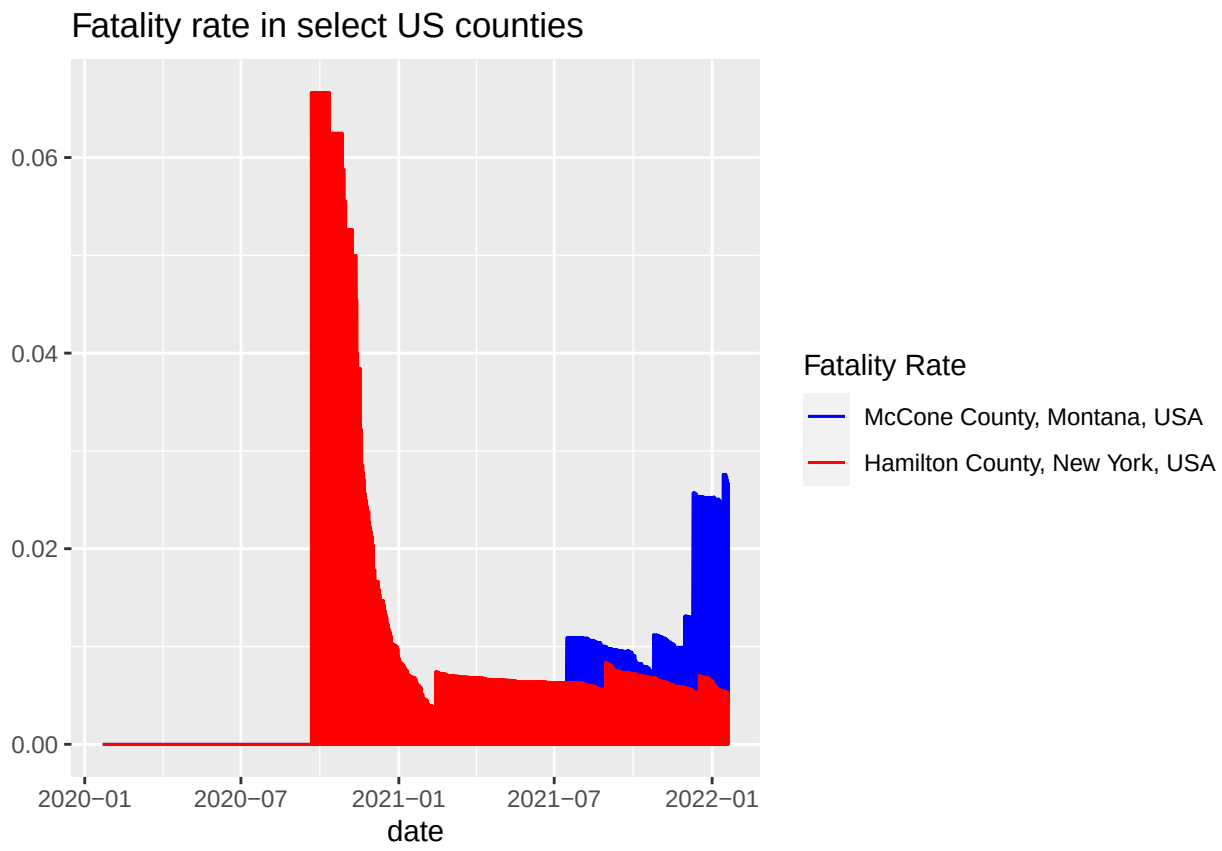
There is little doubt that the mRNA vaccines have proven very successful, particularly against the original variant of SARS-CoV-2 but also to a lesser degree against later variants such as Delta and Omicron. Skepticism against the vaccines have caused many to not get vaccinated even in countries where the vaccines are easily available and vaccination is encouraged. In the diagram below is shown the fatality rate for one of the most vaccinated counties in USA (Hamilton County, New York) against one of the least vaccinated counties (McCone, Montana). According to https://www.nytimes.com/interactive/2020/us/covid-19-vaccine-doses.html McCone county, Montana has 17% fully vaccinated people aged 12+ whereas Hamilton County, New York has 82% fully vaccinated age 12+. Hamilton County experiences the first fatality very early after only 15 registered cases which causes the fatality rate to spike early and then plateau. McCone County on the other hand has a continually growing fatality rate, likely at least part due to lack of support for the vaccination effort. It is worth mentioning that while this seems to prove that the vaccines are effective, if not indeed

essential, the counties are small and a number of other factors such as population demographics, random events etc. could have a huge impact.

```
mccone <- us %>%
  filter(Combined_Key == 'McCone, Montana, US') %>%
  mutate(mccone_fatality = deaths / cases)

hamilton <- us %>% filter(Combined_Key == 'Hamilton, New York, US') %>%
  mutate(hamilton_fatality = deaths / cases)

counties <- mccone %>%
  full_join(hamilton) %>%
  pivot_wider()
```

```
## Joining, by = c("Combined_Key", "date", "cases", "deaths")
```

```
counties <- replace(counties, is.na(counties), 0)

counties %>%
   ggplot(aes(x = date)) +
   geom_line(aes(y=mccone_fatality,color="McCone County, Montana, USA")) +
   geom_line(aes(y=hamilton_fatality,color="Hamilton County, New York, USA")) +
   scale_color_manual(name = "Fatality Rate",
                      values = c("McCone County, Montana, USA" = "blue",
                                 "Hamilton County, New York, USA" = "red")) +
  labs(title = "Fatality rate in select US counties", y=NULL)
```



Data used

confirmed glocal cases from https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_series/:

```
global_cases
```

```
## # A tibble: 204,400 x 4
##    `Province/State` `Country/Region` date    cases
##    <chr>            <chr>            <chr>   <dbl>
##  1 <NA>             Afghanistan      1/22/20     0
##  2 <NA>             Afghanistan      1/23/20     0
##  3 <NA>             Afghanistan      1/24/20     0
##  4 <NA>             Afghanistan      1/25/20     0
##  5 <NA>             Afghanistan      1/26/20     0
##  6 <NA>             Afghanistan      1/27/20     0
##  7 <NA>             Afghanistan      1/28/20     0
##  8 <NA>             Afghanistan      1/29/20     0
##  9 <NA>             Afghanistan      1/30/20     0
## 10 <NA>             Afghanistan      1/31/20     0
## # ... with 204,390 more rows
```

```
summary(global_cases)
```

```
##  Province/State     Country/Region         date               cases
##  Length:204400      Length:204400      Length:204400      Min.   :       0
##  Class :character   Class :character   Class :character   1st Qu.:     201
##  Mode  :character   Mode  :character   Mode  :character   Median :    4075
##                                                           Mean   :  396335
##                                                           3rd Qu.:   81812
##                                                           Max.   :69308111
```

Global deaths from https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_series/:

```
global_deaths
```

```
## # A tibble: 204,400 x 4
##    `Province/State` `Country/Region` date    deaths
##    <chr>            <chr>            <chr>    <dbl>
##  1 <NA>             Afghanistan      1/22/20      0
##  2 <NA>             Afghanistan      1/23/20      0
##  3 <NA>             Afghanistan      1/24/20      0
##  4 <NA>             Afghanistan      1/25/20      0
##  5 <NA>             Afghanistan      1/26/20      0
##  6 <NA>             Afghanistan      1/27/20      0
##  7 <NA>             Afghanistan      1/28/20      0
##  8 <NA>             Afghanistan      1/29/20      0
##  9 <NA>             Afghanistan      1/30/20      0
## 10 <NA>             Afghanistan      1/31/20      0
## # ... with 204,390 more rows
```

```
summary(global_deaths)
```

```
##  Province/State     Country/Region         date               deaths
##  Length:204400      Length:204400      Length:204400      Min.   :     0
##  Class :character   Class :character   Class :character   1st Qu.:     2
##  Mode  :character   Mode  :character   Mode  :character   Median :    59
##                                                           Mean   :  8568
##                                                           3rd Qu.:  1393
##                                                           Max.   :860247
```

confirmed US cases from https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_series/:

```
us_cases
```

```
## # A tibble: 2,439,660 x 3
##    Combined_Key         date    cases
##    <chr>                <chr>   <dbl>
##  1 Autauga, Alabama, US 1/22/20     0
##  2 Autauga, Alabama, US 1/23/20     0
##  3 Autauga, Alabama, US 1/24/20     0
##  4 Autauga, Alabama, US 1/25/20     0
##  5 Autauga, Alabama, US 1/26/20     0
##  6 Autauga, Alabama, US 1/27/20     0
##  7 Autauga, Alabama, US 1/28/20     0
##  8 Autauga, Alabama, US 1/29/20     0
##  9 Autauga, Alabama, US 1/30/20     0
## 10 Autauga, Alabama, US 1/31/20     0
## # ... with 2,439,650 more rows
```

```
summary(us_cases)
```

```
##  Combined_Key          date               cases
##  Length:2439660     Length:2439660     Min.   :      0
##  Class :character   Class :character   1st Qu.:     67
##  Mode  :character   Mode  :character   Median :    942
##                                        Mean   :   6649
##                                        3rd Qu.:   3766
##                                        Max.   :2385721
```

US deaths from https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_series/:

```
us_deaths
```

```
## # A tibble: 2,439,660 x 3
##    Combined_Key         date    deaths
##    <chr>                <chr>    <dbl>
##  1 Autauga, Alabama, US 1/22/20      0
##  2 Autauga, Alabama, US 1/23/20      0
##  3 Autauga, Alabama, US 1/24/20      0
```

```
##  4 Autauga, Alabama, US 1/25/20      0
##  5 Autauga, Alabama, US 1/26/20      0
##  6 Autauga, Alabama, US 1/27/20      0
##  7 Autauga, Alabama, US 1/28/20      0
##  8 Autauga, Alabama, US 1/29/20      0
##  9 Autauga, Alabama, US 1/30/20      0
## 10 Autauga, Alabama, US 1/31/20      0
## # ... with 2,439,650 more rows
```

```
summary(us_deaths)
```

```
##  Combined_Key           date              deaths
##  Length:2439660     Length:2439660     Min.   :    0.0
##  Class :character   Class :character   1st Qu.:    1.0
##  Mode  :character   Mode  :character   Median :   16.0
##                                        Mean   :  118.1
##                                        3rd Qu.:   69.0
##                                        Max.   :28282.0
```

Global recovered from https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_series/:

```
global_recovered
```

```
## # A tibble: 193,450 x 4
##    `Province/State` `Country/Region` date    recovered
##    <chr>            <chr>            <chr>       <dbl>
##  1 <NA>             Afghanistan      1/22/20         0
##  2 <NA>             Afghanistan      1/23/20         0
##  3 <NA>             Afghanistan      1/24/20         0
##  4 <NA>             Afghanistan      1/25/20         0
##  5 <NA>             Afghanistan      1/26/20         0
##  6 <NA>             Afghanistan      1/27/20         0
##  7 <NA>             Afghanistan      1/28/20         0
##  8 <NA>             Afghanistan      1/29/20         0
##  9 <NA>             Afghanistan      1/30/20         0
## 10 <NA>             Afghanistan      1/31/20         0
## # ... with 193,440 more rows
```

```
summary(global_recovered)
```

```
##  Province/State     Country/Region         date             recovered
##  Length:193450      Length:193450      Length:193450      Min.   :       0
##  Class :character   Class :character   Class :character   1st Qu.:       0
##  Mode  :character   Mode  :character   Mode  :character   Median :     218
##                                                           Mean   :  121434
##                                                           3rd Qu.:    9218
##                                                           Max.   :30974748
```

Final thoughts

Although the datasets used are regarded as some of the best there are large differences in the reporting from one country to another. China, India and Peru are some of those countries that are said to be under-reporting either purposefully or due to lack of resources, compounded by the fact that a lot of COVID19 cases have never been identified due to infections being asymptomatic or only with mild symptoms. Also many countries have had technical issues causing data to be not entirely faithful, Denmark for example have had days where reporting was not possible resulting in a surge of reported cases when the systems were working again. Finally there has been much discussion on what the counts should include (should death counts include anyone who were infected at the time of death or only those where COVID19 was proven to be the primary cause of death). All of this means that the data presented is not entirely truthful, but it is as close as it can be at the time of writing. My personal aim of this report has been to highlight a few of the success stories in this long struggle against SARS-CoV-2. Obviously there have been many more situations where it is not possible to find a positive angle which I could also have highlighted but I felt that that aspect has already been well covered.

It has been an extremely interesting subject to dive into and one that would probably be large and complex enough for a lifetime of study. With a few diagrams it is able to follow the effects of politics and new variants of the virus, as well as development of vaccines and treatments. In particular I found it heartening to see what appears the first signs of the virus becoming less lethal; a promise of better times to come.