COVID Data

C. Godsey

2023-04-01

Introduction

We're going to explore the current COVID-19 data for the purpose of uncovering any interesting trends. Our focus for this article will be on the change in mortality over time. Much of the focus early on in COVID was on controlling transfer and new cases, but regardless of the number of new cases at any point, there should be important data on the mortality associated with however many cases there are.

Covid 19 Data

```
#url_base <- 'https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_ti
#files <- c("time_series_covid19_confirmed_US.csv",  "time_series_covid19_confirmed_global.csv", "time_
#urls <- str_c(url_base, files)
#US_cases_raw <- read_csv(urls[1])
#global_cases_raw <- read_csv(urls[2])
#US_deaths_raw <- read_csv(urls[2])
#global_deaths_raw <- read_csv(urls[3])

global_deaths_raw <- read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_co
global_cases_raw <- read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_cov
```

Tidy and Preprocess the Data

We transform the columnar date data to provide total number of cases and deaths per area.

```
global_cases <- global_cases_raw %>%
  select(-c(Lat, Long)) %>%
  rename(Province_State = `Province/State`, Country_Region = `Country/Region`) %>%
  pivot_longer(cols = -c(Province_State, Country_Region),
               names_to = 'Date', values_to = 'Cases') %>%
  mutate(Date = mdy(Date))

global_deaths <- global_deaths_raw %>%
  select(-c(Lat, Long)) %>%
  rename(Province_State = `Province/State`, Country_Region = `Country/Region`) %>%
  pivot_longer(cols = -c(Province_State, Country_Region),
               names_to = 'Date', values_to = 'Deaths') %>%
  mutate(Date = mdy(Date))
```

Merge the Data and Engineer our Fatality Value

Now we want to merge the disparate cases and deaths datasets to a single dataset. The average time to death since COVID infection is 3 weeks. We offset the case dates by 21 days so we can more accurately compute a fatality per case ratio, as we're interested in the deaths related to roughly the time when COVID was contracted. With these merged datasets, we're able to computer the death/case (fatality) ratio. We are filtering out just the major geographic areas, to help account for a situation we'll explain in the next graph.

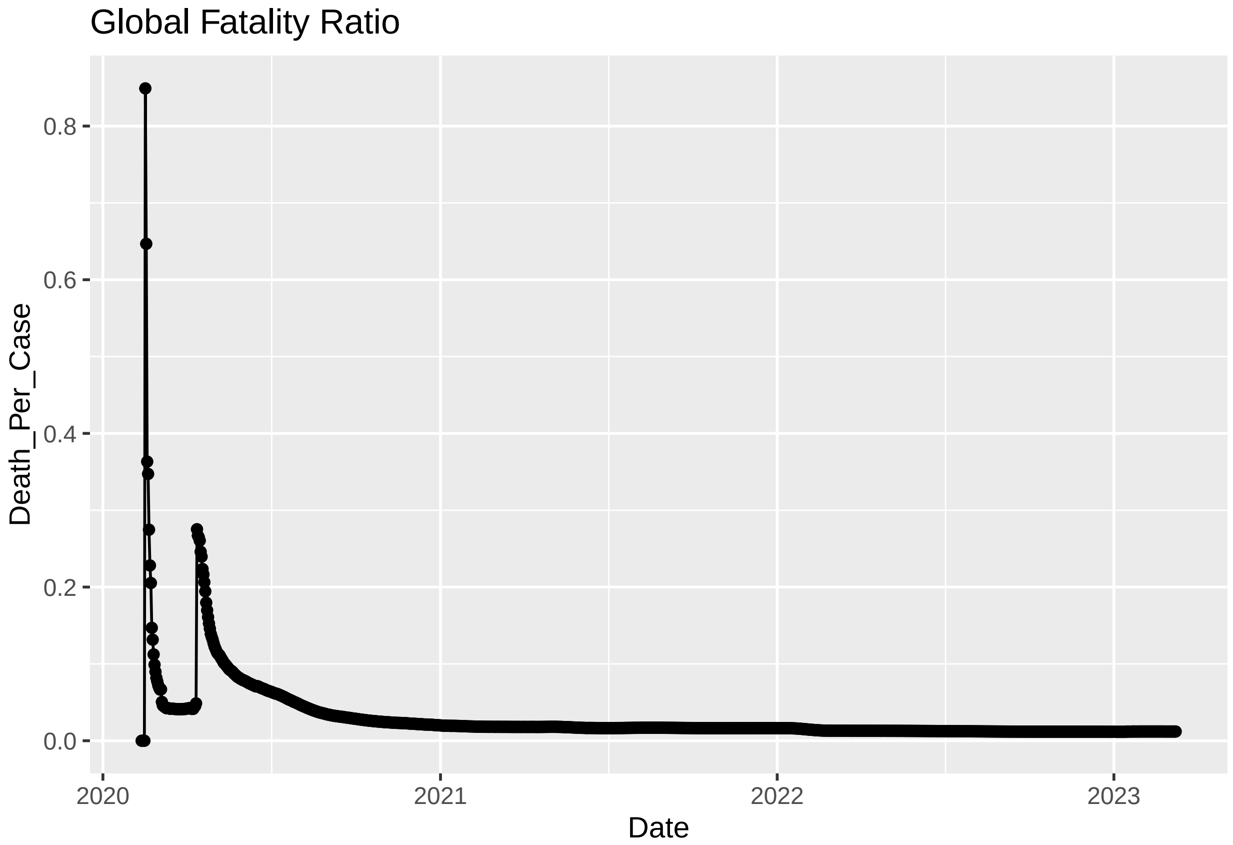
```r
global <- global_cases %>%
  mutate(Date = Date + days(21)) %>%
  full_join(global_deaths) %>%
  filter(Country_Region %in% c('US', 'India', 'China', 'Indonesia')) %>%
  summarise(.by = c(Country_Region, Date), Cases = sum(Cases), Deaths = sum(Deaths)) %>%
  mutate(Death_Per_Case = Deaths/Cases) %>%
  filter(Death_Per_Case < 1)
global[sapply(global, is.infinite)] <- NA

summary(global)
```

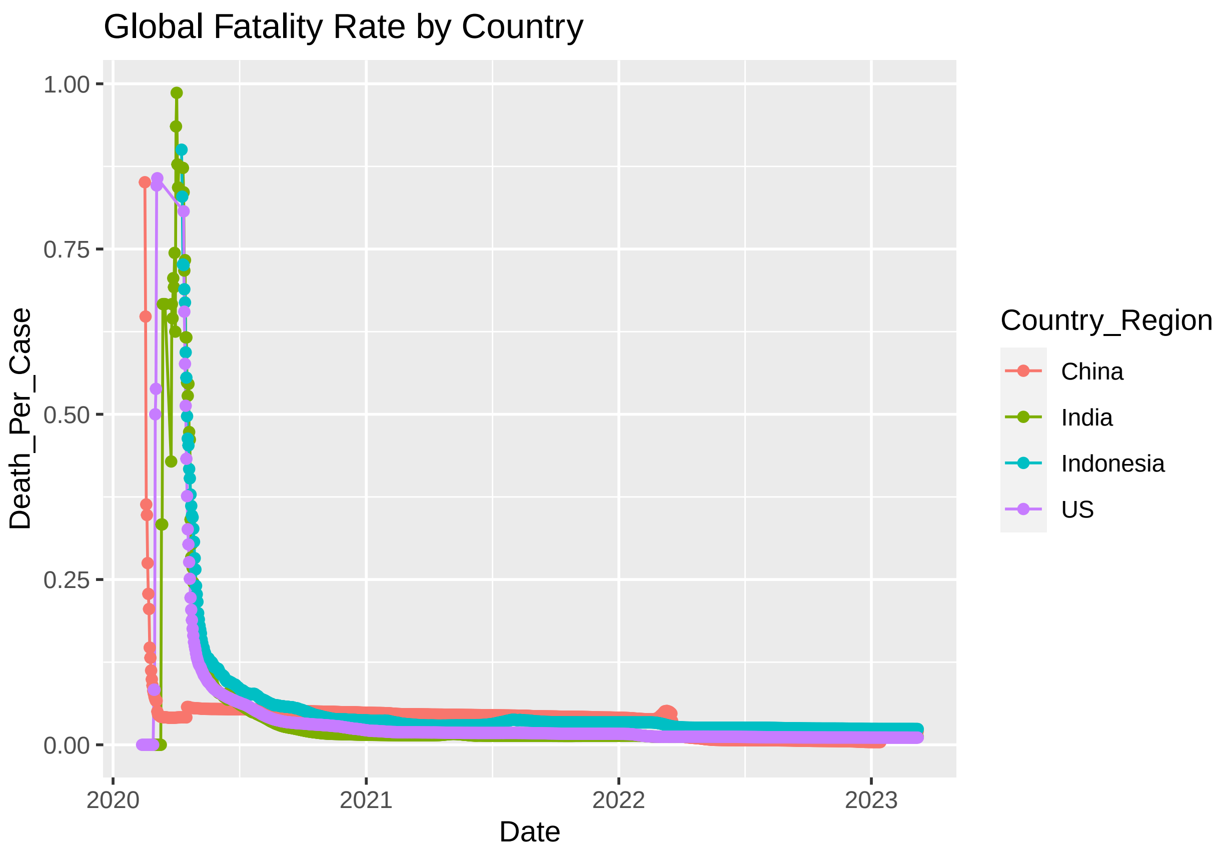
```
##  Country_Region          Date                Cases               Deaths
##  Length:4369        Min.   :2020-02-12   Min.   :        1   Min.   :      0
##  Class  :character  1st Qu.:2020-12-10   1st Qu.:   232628   1st Qu.:  12431
##  Mode   :character  Median :2021-09-09   Median :  4903515   Median : 144220
##                     Mean   :2021-09-08   Mean   : 19327574   Mean   : 272937
##                     3rd Qu.:2022-06-09   3rd Qu.: 33845569   3rd Qu.: 516479
##                     Max.   :2023-03-09   Max.   :103083910   Max.   :1123836
##  Death_Per_Case
##  Min.   :0.00000
##  1st Qu.:0.01258
##  Median :0.02415
##  Mean   :0.03866
##  3rd Qu.:0.04203
##  Max.   :0.98630
```

Exploration

We started by first exploring the global fatality rate. As you can see in the following graph, there are two peaks in fatality. There are many possible causes for this, but we'll showcase the source in the next graph.



So as we can see, there is definitely an exponential decrease in mortality rate over time. But as you might notice, we have these two peaks. With the obvious decrease, having a natural second spike in mortality does not make a lot of sense, and the data does not look continuous. We do know, historically, that COVID did not start to spread in every area at the same time. Each country had their own initial onset of infection, so lets take a look at each of these 4 most popular countries separately.

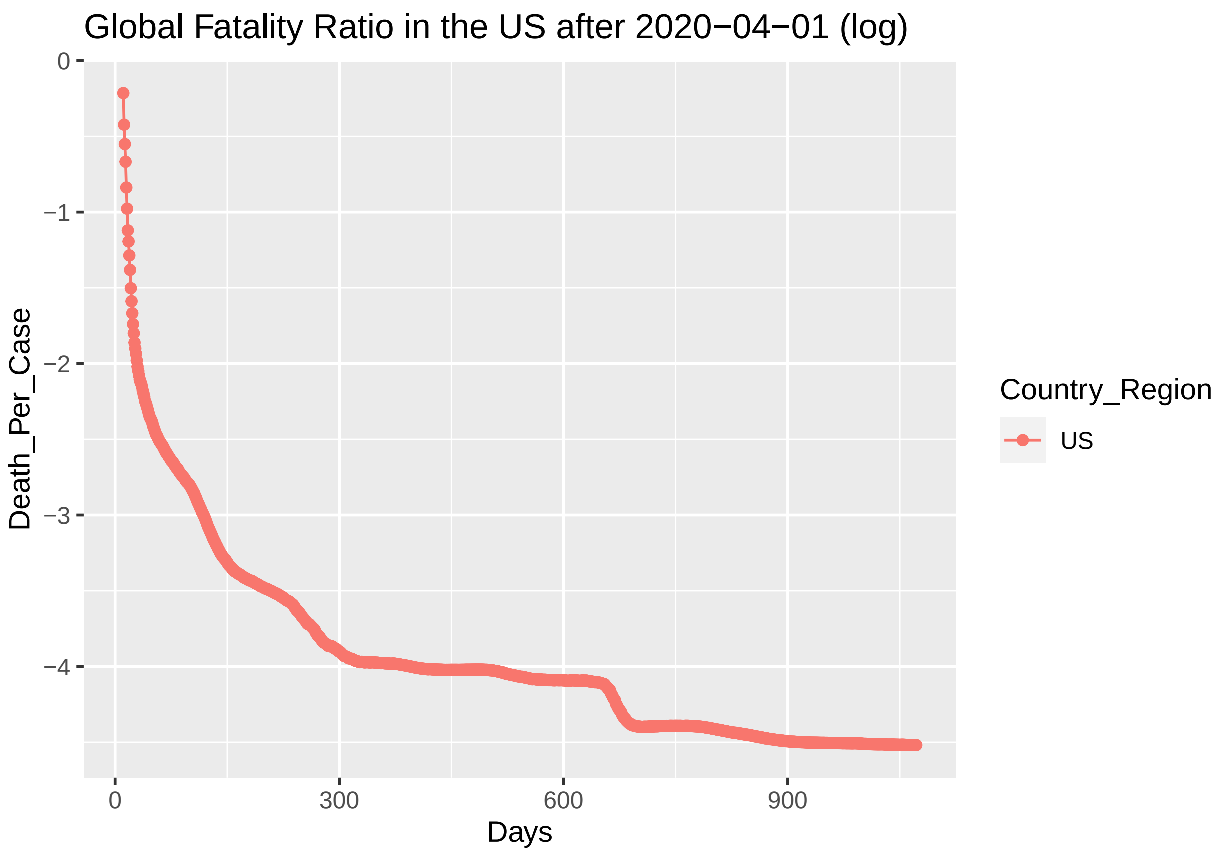


When looking at the data for these 4 countries separately, we can see a much more reasonable pattern. The decrease in mortality is in fact exponential, and the two spikes we saw in the first graph are just due to the time in which each country had their initial outbreaks. So accounting for that, we can see a very clear and continuous drop in fatality over time, and the drop in fatality is similar in relation to the date of the initial outbreak.

While the decay in fatality is similar in relation, there are some subtle differences in the decay. So next, we're going to take a look at modeling each country independently. The US has a large population and has been rather diligent about reporting data, so we'll model the US.

The US Fatality

Let's look at the US data, and we'll trim off the first few months of data to get a cleaner visual. We also plot the graph logarithmically so we can better view the quality beyond the exponential nature of the data.

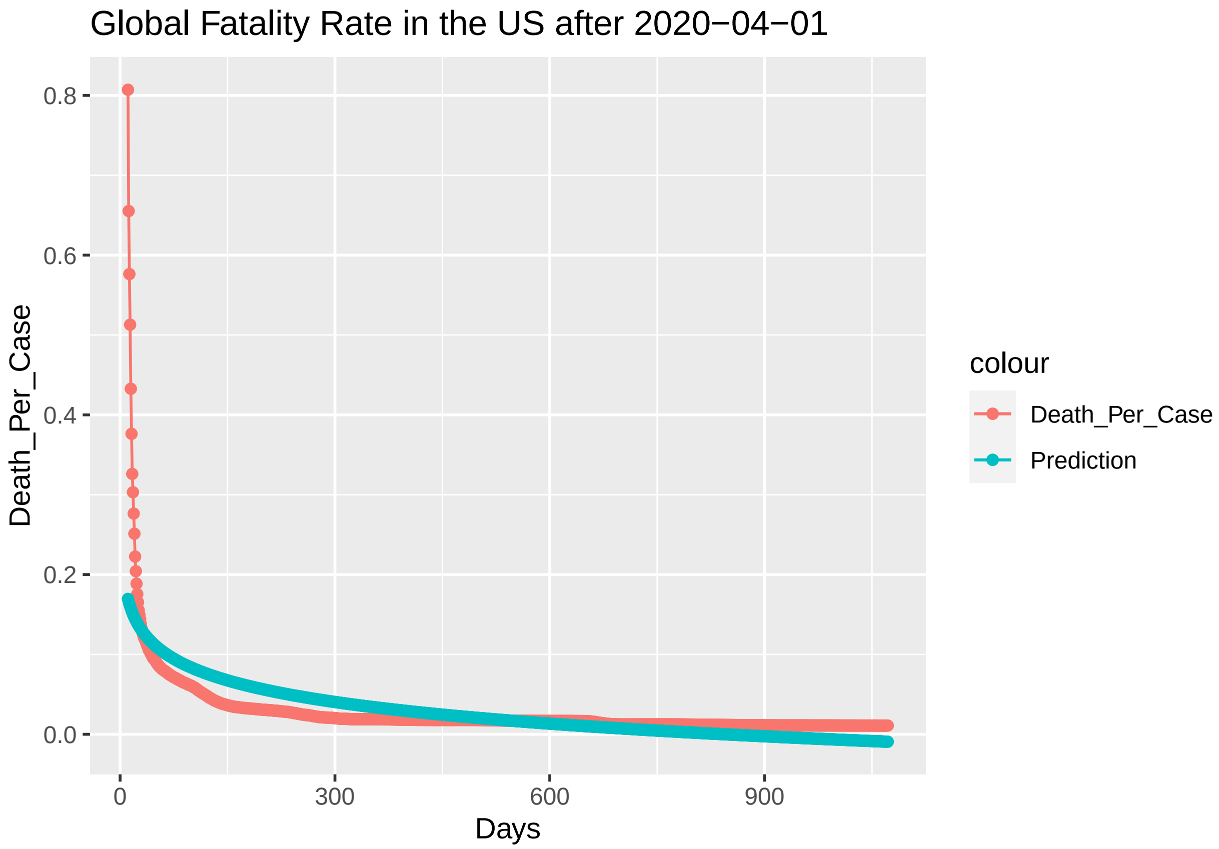


The data does have some qualities outside of the simple logarithmic property. The remaining characteristics do not seem continuous, so let's just create a logarithmic model of the data.

```
model <- lm(data=filtered_us_data, Death_Per_Case ~ log(Days))
summary(model)
```

```
##
## Call:
## lm(formula = Death_Per_Case ~ log(Days), data = filtered_us_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.03126 -0.01920  0.00090  0.01093  0.63754
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.263209   0.007728   34.06   <2e-16 ***
## log(Days)   -0.039057   0.001269  -30.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03668 on 1060 degrees of freedom
## Multiple R-squared:  0.4718, Adjusted R-squared:  0.4713
## F-statistic: 946.8 on 1 and 1060 DF,  p-value: < 2.2e-16
```

```
filtered_us_data %>%
  mutate(Prediction = predict(model)) %>%
  ggplot(aes(x = Days)) +
  geom_line(aes(y = Death_Per_Case, color="Death_Per_Case")) +
  geom_point(aes(y = Death_Per_Case, color="Death_Per_Case")) +
  geom_line(aes(y = Prediction, color="Prediction")) +
  geom_point(aes(y = Prediction, color="Prediction")) +
  labs(title = "Global Fatality Rate in the US after 2020-04-01")
```



The model seems to roughly model the logarithmic nature in the data, but there some specific nuances that are not reflected in the general model. This makes sense as this pandemic was punctuated by several discrete events that could have had a large impact on mortality, such as: immunity, vaccination, lockdowns, and changes in treatment.

Conclusion

In this article, we explored an area of the COVID data that is not often explored. It shows the surprising ability of a human population to adapt to a major outbreak. Advances in medical technology and the increase in immunity can allow a population to adapt and overcome fatality even as the diseases grows and spreads further.

Bias

The article was produced with a conscious awareness of potential bias. We have avoided drawing any firm conclusions as to why the fatality rate dropped over time and instead chose to identify the trend itself and point to some possible causal elements. There is also a chance that this trend reflects changes in testing and infection identification, and may not truly reflect the changes in mortality.

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.2 LTS
## 
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so
## 
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C              
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8    
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8   
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                 
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C            
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C       
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] lubridate_1.9.2 forcats_1.0.0   stringr_1.5.0   dplyr_1.1.1    
##  [5] purrr_1.0.1     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1   
##  [9] ggplot2_3.4.1   tidyverse_1.3.1
## 
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.0 xfun_0.38        haven_2.5.2      colorspace_2.1-0
##  [5] vctrs_0.6.1      generics_0.1.3   htmltools_0.5.5  yaml_2.3.7      
##  [9] utf8_1.2.3       rlang_1.1.0      pillar_1.9.0     glue_1.6.2      
## [13] withr_2.5.0      DBI_1.1.3        bit64_4.0.5      dbplyr_2.3.2    
## [17] modelr_0.1.11    readxl_1.4.2     lifecycle_1.0.3  munsell_0.5.0   
## [21] gtable_0.3.3     cellranger_1.1.0 rvest_1.0.3      evaluate_0.20   
## [25] labeling_0.4.2   knitr_1.42       tzdb_0.3.0       fastmap_1.1.1   
## [29] curl_5.0.0       parallel_4.1.2   fansi_1.0.4      highr_0.10      
## [33] broom_1.0.4      scales_1.2.1     backports_1.4.1  vroom_1.6.1     
## [37] jsonlite_1.8.4   farver_2.1.1     bit_4.0.5        fs_1.6.1        
## [41] hms_1.1.3        digest_0.6.31    stringi_1.7.12   grid_4.1.2      
## [45] cli_3.6.1        tools_4.1.2      magrittr_2.0.3   crayon_1.5.2    
## [49] pkgconfig_2.0.3  xml2_1.3.3       reprex_2.0.2     timechange_0.2.0
## [53] rmarkdown_2.21   httr_1.4.5       rstudioapi_0.14  R6_2.5.1        
## [57] compiler_4.1.2
```