COVID_universal_healthcare_data_analysis

D. Fritts

2025-09-01

Introduction

The central question I will attempt to answer in this analysis is **"Does access to universal healthcare affect the COVID case to death ratio?"**. Universal healthcare is often cited as one of the reasons for the success or failure of a country's health, depending on your political bias. Since COVID was the most significant health crisis of our lives to date, it seems like an interesting question to see if universal healthcare helped countries deal with COVID or not. The way I am measuring this is the "case to death ratio". The case to death ratio can be described as: "of the people who contracted COVID, what percent of them passed away?".

Sources We will be obtaining the COVID cases and deaths data from the John Hopkins COVID data repository, located here: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series. The two data sets we are obtaining for the analysis are: "time_series_covid19_deaths_global.csv" for the deaths and "time_series_covid19_confirmed_global.csv" for the cases.

Our John Hopkins data set does not include information about which countries do or do not have universal healthcare, so I have obtained it from a third party source: https://worldpopulationreview.com/country-rankings/countries-with-universal-healthcare. I have included this data set as a csv in the repository, called "countries-with-universal-healthcare-2025.csv". Please make sure this file is in the same directory as the rmd file for proper knitting.

Data Load and Setup

Setting up environment variables and loading in the relevant libraries.

```
# Loading in necessary libraries
library(tidyverse)
```

Loading in the two John Hopkins data sets. The first one details the global COVID case numbers for each country on each day. The other details the global deaths attributed to COVID for each country on each day. The time period this covers is 2020 to 2023.

```
# For this analysis I am only going to work with the global data.

url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi

file_names <- c(
  "time_series_covid19_confirmed_global.csv",
  "time_series_covid19_deaths_global.csv")
```

```
urls <- str_c(url_in, file_names)

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

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Loading in the universal healthcare data from worldpopulationreview, detailing which countries have universal healthcare and which do not as of 2025.

```
healthcare <- read_csv("./countries-with-universal-healthcare-2025.csv")
```

```
## Rows: 78 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): flagCode, country, CountriesWithUniversalHealthcare
## dbl (1): CountriesWithUniversalHealthcareUniversalHealthcareIndex2021
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Preparation

In this step, the COVID datasets are cleaned and transformed. Both the cases and deaths data sets come in wide form, meaning that there is a column for each day recorded that denotes the number of deaths or cases on that given day. I will be pivoting those down to long form, meaning that there are two columns: "cases" and "deaths" that represents the occurence of deaths and cases on that day for a given country.

These two datasets, cases and deaths are joined together, into a data set called "global". Afterwords, the data set containing information about if a country has universal healthcare or not is joined to global to create "global_hc".

Columns are renamed for clarity, and column data types are standardized away from character types.

```
# Pivoting out global cases and removing Lat and Long
global_cases_tidy <- global_cases %>% pivot_longer(cols = -c('Province/State', 'Country/Region', 'Lat',

# Pivoting out global deaths and removing Lat and Long
global_deaths_tidy <- global_deaths %>% pivot_longer(cols = -c('Province/State', 'Country/Region', 'Lat

# Joining global cases and deaths, cleaning up column names and changing date(string) to date (date)
global <- global_cases_tidy %>%
  full_join(global_deaths_tidy) %>%
  rename(Country_Region = 'Country/Region',
         Province_State = 'Province/State') %>%
  mutate(date = mdy(date))
```

```
## Joining with `by = join_by(`Province/State`, `Country/Region`, date)`
```

```
# Joining universal healthcare data into the covid dataset.
global_hc <- global %>%
  left_join(select(healthcare, country, CountriesWithUniversalHealthcare),
            by = c("Country_Region" = "country")) %>% select(-c("Province_State")) %>%
  mutate(Country_Region = as.factor(Country_Region),
         CountriesWithUniversalHealthcare = replace_na(CountriesWithUniversalHealthcare, "No"),
         CountriesWithUniversalHealthcare = as.factor(CountriesWithUniversalHealthcare)) %>%
  rename(has_uhc = CountriesWithUniversalHealthcare, country = Country_Region)

summary(global_hc)
```

```
##            country            date                cases
##  China         : 38862   Min.   :2020-01-22   Min.   :        0
##  Canada        : 18288   1st Qu.:2020-11-02   1st Qu.:      680
##  United Kingdom: 17145   Median :2021-08-15   Median :    14429
##  France        : 13716   Mean   :2021-08-15   Mean   :   959384
##  Australia     :  9144   3rd Qu.:2022-05-28   3rd Qu.:   228517
##  Netherlands   :  5715   Max.   :2023-03-09   Max.   :103802702
##  (Other)       :227457
##      deaths         has_uhc
##  Min.   :      0   No :154305
##  1st Qu.:      3   Yes:176022
##  Median :    150
##  Mean   :  13380
##  3rd Qu.:   3032
##  Max.   :1123836
##
```

Here, we are deriving the variable we are concerned with: death_to_case_ratio. Death_to_case_ratio is defined by total_deaths / total_cases for a country and given year. A higher death_to_case_ratio indicates that a country had a higher occurence of death for those who contracted COVID, and the inverse is true for lower death_to_case_ratios.

```
yearly_ratios <- global_hc %>%
  filter(cases > 0 & deaths > 0) %>%
  group_by(country, year = year(date)) %>%
  summarise(total_cases = max(cases),
```

```
            total_deaths = max(deaths),
            has_uhc = first(has_uhc)) %>%
  mutate(death_to_case_ratio = total_deaths / total_cases) %>%
  arrange(death_to_case_ratio) %>%
  filter(death_to_case_ratio <= 1)
```

```
## `summarise()` has grouped output by 'country'. You can override using the
## `.groups` argument.
```

```
summary(yearly_ratios)
```

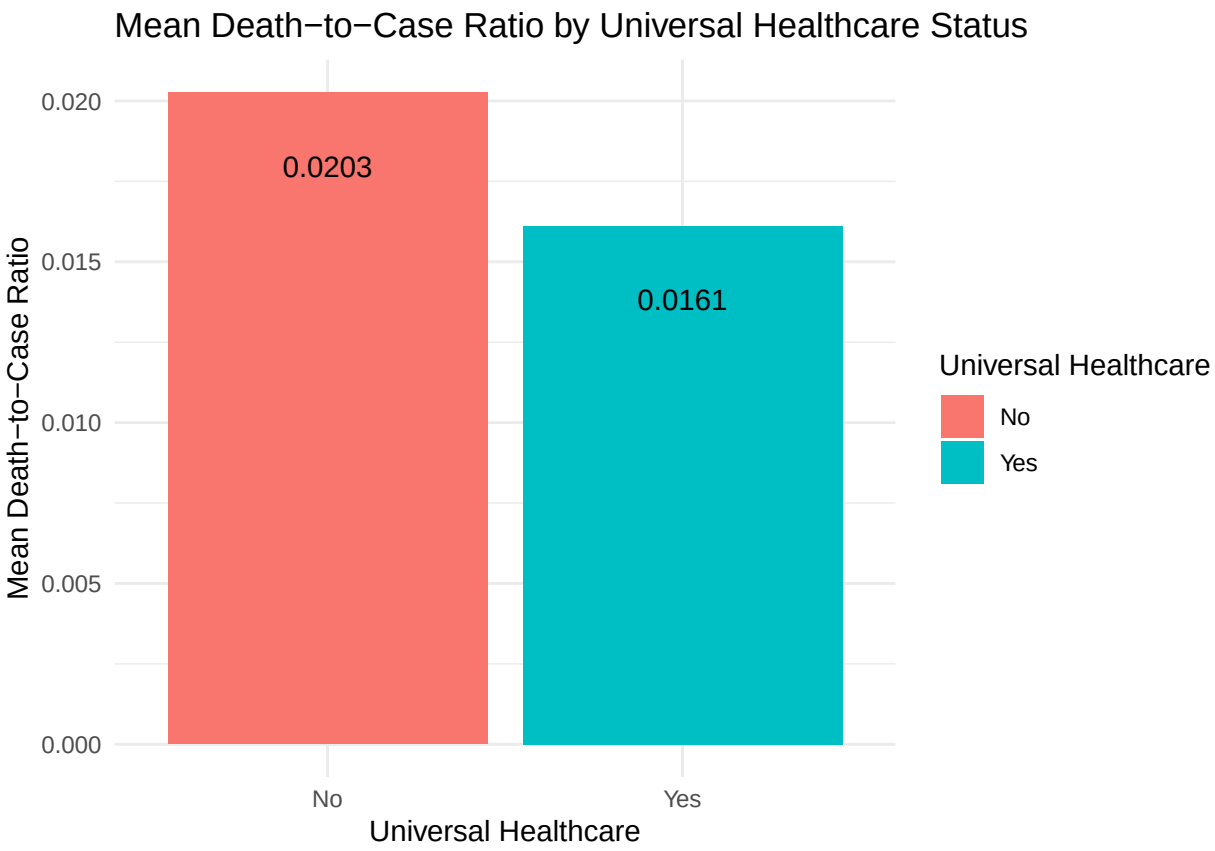
```
##                 country         year       total_cases
##  Afghanistan        :  4   Min.   :2020   Min.   :        7
##  Albania            :  4   1st Qu.:2021   1st Qu.:    24825
##  Algeria            :  4   Median :2022   Median :   201785
##  Andorra            :  4   Mean   :2022   Mean   :  2226603
##  Angola             :  4   3rd Qu.:2023   3rd Qu.:  1083377
##  Antigua and Barbuda:  4   Max.   :2023   Max.   :103802702
##  (Other)            :730
##   total_deaths       has_uhc   death_to_case_ratio
##  Min.   :      1.0   No :492   Min.   :0.0001906
##  1st Qu.:    294.2   Yes:262   1st Qu.:0.0073562
##  Median :   2526.5             Median :0.0138431
##  Mean   :  27571.8             Mean   :0.0188263
##  3rd Qu.:  13975.5             3rd Qu.:0.0222132
##  Max.   :1123836.0             Max.   :0.2906146
##
```

Initial Analysis and Visualization

Overall Mean Ratios First I want to examine the overall trend. On average, how do countries who have universal healthcare compare to countries without universal healthcare in terms of death to case ratio when when looking at all the recorded years?

```
# Summarize mean ratios
overall_ratios_summary <- yearly_ratios %>%
  group_by(has_uhc) %>%
  summarise(mean_ratio = mean(death_to_case_ratio, na.rm = TRUE))

# Plot
ggplot(overall_ratios_summary, aes(x = has_uhc, y = mean_ratio, fill = has_uhc)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = round(mean_ratio, 4), vjust=4)) +
  labs(title = "Mean Death-to-Case Ratio by Universal Healthcare Status",
       x = "Universal Healthcare",
       y = "Mean Death-to-Case Ratio",
       fill = "Universal Healthcare") +
  theme_minimal()
```




```
overall_ratios_summary
```

```
## # A tibble: 2 x 2
##   has_uhc mean_ratio
##   <fct>        <dbl>
## 1 No          0.0203
## 2 Yes         0.0161
```

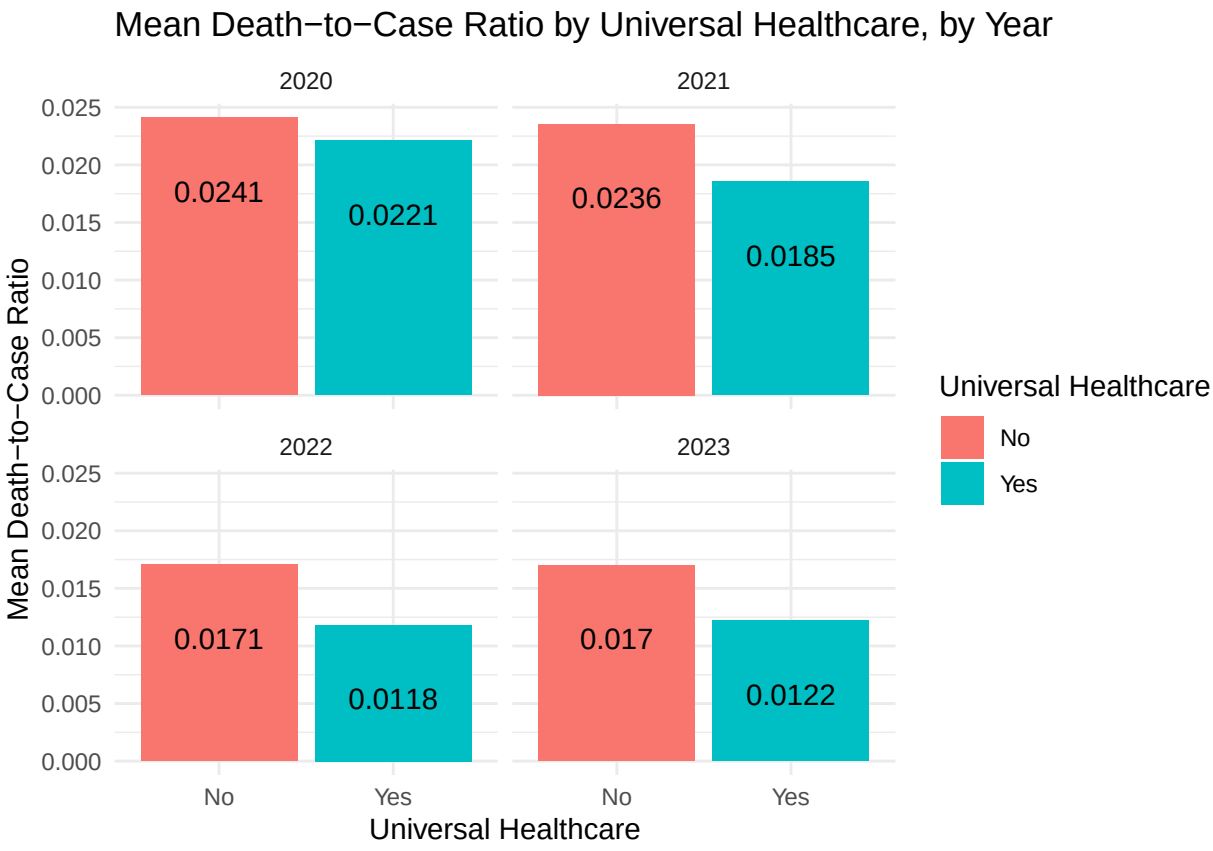
Looking at the graph and summary, on average, it looks like countries with universal healthcare have lower death_to_case_ratios, meaning that their citizens had a lower chance of dying from COVID once they contracted it. However, that overall difference in mean ratios is small, only a difference of ~0.0042. This is just a high level view, so lets dive deeper to obtain a greater understanding of what's going on here.

Yearly Mean Ratios Here, we break out the data by year and show the difference in death to case ratios.

```
# Summarize mean ratios by has_uhc and year
yearly_ratios_summary <- yearly_ratios %>%
  group_by(has_uhc, year) %>%
  summarise(mean_ratio = mean(death_to_case_ratio, na.rm = TRUE), .groups = "drop")

# Faceted bar plot
ggplot(yearly_ratios_summary, aes(x = has_uhc, y = mean_ratio, fill = has_uhc)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ year) +
```

```
  geom_text(aes(label = round(mean_ratio, 4), vjust=4)) +
  labs(title = "Mean Death-to-Case Ratio by Universal Healthcare, by Year",
       x = "Universal Healthcare",
       y = "Mean Death-to-Case Ratio",
       fill = "Universal Healthcare") +
  theme_minimal()
```




```
yearly_ratios_summary
```

```
## # A tibble: 8 x 3
##   has_uhc  year mean_ratio
##   <fct>   <dbl>      <dbl>
## 1 No       2020     0.0241
## 2 No       2021     0.0236
## 3 No       2022     0.0171
## 4 No       2023     0.0170
## 5 Yes      2020     0.0221
## 6 Yes      2021     0.0185
## 7 Yes      2022     0.0118
## 8 Yes      2023     0.0122
```

Breaking out the data by year, we can see that the overall trend still holds: countries with universal healthcare had lower rates of dying from COVID than countries without universal healthcare across all years. Interestingly, it seems in the initial wave of the virus, 2020, the difference was smallest, but as time progressed the gap between those with universal healthcare and those without universal healthcare generally

grew. The only exception to this trend was 2024 in which they stayed roughly the same. Put another way, the death_to_case_ratio fell faster over time in countries with universal healthcare compared to those without universal healthcare. However, the variation between them is small, so I want to dig deeper in to see if these differences hold statistical significance.

Wilcoxon Test: Comparing Distributions

Since we are interested in comparing distributions, I am going to use the Wilcoxon Rank-Sum test. This allows us to check if two independent groups have the same distribution / median. In this case, the two groups are countries with universal healthcare and those without universal healthcare. The distributions in question are the death_to_case_ratios for each group and for each year. The null hypothesis in this test asserts that the distrubtions between the two groups are the same, meaning there is no significant difference between the two groups. If the analysis reveals a p-value <0.05, there is good reason to reject this null hypothesis.

```
# Wilcoxon test for each year
p_values <- yearly_ratios %>%
  group_by(year) %>%
  summarise(p_value = wilcox.test(death_to_case_ratio ~ has_uhc)$p.value,
            .groups = "drop")

# Print results
print(p_values)
```

```
## # A tibble: 4 x 2
##    year p_value
##   <dbl>   <dbl>
## 1  2020  0.644
## 2  2021  0.0973
## 3  2022  0.0489
## 4  2023  0.0951
```

As you can see, the p-values for each year differ significantly. In 2020, it seems that universal healthcare did not differentiate the two groups significantly - if you will refer back to our yearly graph you can see that this is the year in which the gap between the two was smallest. 2022, on the other hand, does show a p-value <0.05, meaning there is good reason to be beleive that a country having universal healthcare does affect the death to case ratio, at least in a small way. In the case of 2021 and 2023, while the p-value is not less than 0.05, it is quite close, being ~0.09. This, combined with the overall trend, suggests to me that there may be something to this - further analysis is needed.

Linear Model Analysis

Here I will attempt to use a linear model to see if the death_to_case_ratio is a function of a country having universal healthcare and the year.

```
ratios_model <- yearly_ratios %>%
  mutate(has_uhc = as.factor(has_uhc),
         year = as.factor(year))

model <- lm(death_to_case_ratio ~ has_uhc + year, data = ratios_model)
summary(model)
```

```
##
## Call:
## lm(formula = death_to_case_ratio ~ has_uhc + year, data = ratios_model)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.024110 -0.010288 -0.005065  0.003693  0.265667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.024947   0.001990  12.534  < 2e-16 ***
## has_uhcYes  -0.004305   0.001901  -2.265  0.02382 *
## year2021    -0.001639   0.002606  -0.629  0.52959
## year2022    -0.008201   0.002580  -3.178  0.00154 **
## year2023    -0.008110   0.002580  -3.143  0.00174 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02485 on 749 degrees of freedom
## Multiple R-squared:  0.02785,    Adjusted R-squared:  0.02266
## F-statistic: 5.364 on 4 and 749 DF,  p-value: 0.0002911
```

Interpreting the Model Results

Residuals

Residuals represent the difference between the actual values and the predicted values - essentially the error in predictions (Actual - Predicted). The residuals have a relatively wide distribution towards the positive end (max: 0.265667), but are tight around the lower end (min: -0.024110 and median: -0.005065). This suggests that there is outlier country+year which has a really high death_to_case ratio, or perhaps a value that the model significantly undershot.

Coefficients

1. Intercept (0.0249, p-value=2e-16): The intercept, represents the mean death_to_case_ratio for countries without universal healthcare in the year 2020. This is our baseline ratio, of 2.49%, meaning this is the expected death to case ratio for a country without universal healthcare in 2020. This is the starting point to which the other coefficients are compared.
2. has_uhcYes(-0.0043, $p < 0.02382$): This coefficient is significant at $p < 0.05$. This means that countries with universal healthcare have, on average, 0.43% lower death_to_case_ratio values than those countries without universal healthcare. The difference between them is not enormous, but having universal healthcare does have a statistically significant impact on lowering death to case rate.
3. year2021 (-0.0016, $p = 0.5296$): This coefficient is not significant, as it has a p-value of 0.5296. Meaning that while the average rate was 0.0016 lower than 2020, this could be due to chance.
4. year2022 (-0.0082, $p = 0.0015$): This coefficient is significant at $p < 0.01$. Ratios in 2022 were lower than 2020 by 0.00082 - small difference but still an improvemnt.
5. year 2023 (-0.00081, $p = 0.0017$): This coefficient is significant at $p < 0.01$. Just like 2022, ratios were 0.00081 smaller than 2020, a small change but an statistically significant improvement.

Since the p-values for the years start high and then got smaller, the interpretation is that as countries with universal healthcare got better at reducing their death to case rate as time went on, even if it initially did not make an enormous difference.

Residual Standard Error (RSE) = 0.02485

RSE measures the average deviation of the actual death_to_case_ratio from the model's predictions. On average, predictions are off by 2.485%. This indicates further unexplained variability.

R-Squared: Multiple = 0.02785, Adjusted = 0.02266

The R-Square value signifies how much variance is explained in death_to_case_ratios by the model. In this case it is only around 2.5%, meaning that year and whether a country has universal healthcare or not does not explain most of the variation. Other factors such as GPD, region, population size, testing procedures, and vaccination rates likely explain much of the remaining variability.

F-statistics = 5.364 on 4 and 749 DF, p-value = 0.0002911

Since the p-value is so low, this means that the overall model is statistically significant. This means that the predictors of "has_uhc" and "year" together provide better predictions than a model with no predictors. Combining this fact with the low R-squared value means that, while there is something here, the improvement in prediction capability is modest.

749 DF means that my sample size is ~750. More granularity, perhaps breaking it down by month, may reveal more definite trends.

Conclusion

It seems that the presence of universal healthcare is modestly linked to better outcomes for countries in terms of COVID death to case ratios. Ratios improved significantly as years went on, suggesting that there are other factors at play such as vaccination rates, quarantine protocols, virus mutation, and better treatments. Other potential factors which are not included but likely played a significant role are annual GPD, quality of healthcare, numbers of hospitals, population sizes, and geographic regions.

Biases

Data Biases

The universal healthcare dataset contains information about the state of countries in 2025. It is possible that some countries did not have universal healthcare in 2020, but adopted it in the later, more recent years. This is a bias which could skew results. According to my research, no countries have transitioned from a state of not having universal healthcare to having universal healthcare between 2020 and 2025, but if I missed it then that could skew results.

Another data bias could be in the way that deaths and cases were tracked. For the deaths, these presumably were not overall deaths, but rather the deaths attributed to COVID. Improperly attributing the cause of death to COVID, especially between time periods and countries, could heavily bias this data into unreliability. For the cases, testing likely rose, peaked and declined at points during 2020 and 2025, rather than staying stable. More testing likely means more cases. Different countries likely peaked at different points due to policies, severity and political structure. It is these last two sources of bias that, to me, represent the most significant risk in drawing conclusions from this analysis. To mitigate this, one must verify that the data is obtained from a reputable source, which in my opinion it is. Additionally, if you look at the README.md for the covid data, it details where all of their information has come from. Since this data set is aggregating many different sources, mostly from governments, it's important to consider how different countries counted and aggregated deaths and cases, and how these may differ from other countries and other time periods.

Personal Biases

I don't have any particular strong feelings on universal healthcare, but if I did that could possibly skew my perception of the problem, twisting the data with the hopes of finding that universal healthcare did or did

not help in this particular scenario. I don't understand the true implications there would be were I to live in a society with universal healthcare - some say it's the best, others say it's the worst - I don't have any personal experience to relate to it, nor any particularly bad or good experiences with our current healthcare system. As for COVID, I have a pretty negative bias towards the whole situation, seeing many policies as mishandled and unnecessary. This could potentially skew my interpretation. If I was doing an analysis on the deadliness of COVID, this bias would be front and center to any of my conclusions. However, since my conclusion primary comes from: 1) The data shows that universal healthcare likely had a small but significant impact on saving lives 2) Much more data is needed to verify these claims in any certainty and comes directly from data and modelling, I am hopeful that this bias of mine does not color the conclusion in any significant way. Finally, I generally consider it a bad thing when people die unnecessarily, and so have stated that it's a worse outcome when the death to case ratio is higher. Many perspectives might disagree with this, and there are likely plenty of arguments to be made from many angles that death is not a bad thing, or that higher rates of death might be beneficial in XYZ way. This could color any of my statements about what is bad or good, but fundamentally the analysis is on the way universal healthcare affect the ratio, not on whether that effect is a good or bad thing.