Excess Mortality in the Age of COVID-19

Data Mining for Insights into Death

Julie Kirkpatrick
University of Colorado Boulder
Boulder Colorado USA
juki9272@colorado.edu

Alex Melnick
University of Colorado Boulder
Boulder Colorado USA
alme9138@colorado.edu

Kyle Tomlinson
University of Colorado Boulder
Boulder Colorado USA
kyto3309@colorado.edu

ABSTRACT

In wake of the ongoing COVID-19 pandemic, much data has been collected to better and further understand the situation. This data in combination with regularly collected mortality data has the potential to reveal interesting patterns. Our project will investigate the relationship between mortality rates and COVID-19.

CCS CONCEPTS

• Applied computing~Life and medical sciences~Health care information systems

KEYWORDS

COVD-19, Mortality, Data Mining, Excess Mortality, Coronavirus, Pandemic, Data Collection

ACM Reference format:

Julie Kirkpatrick, Alex Melnick and Kyle Tomlinson. Excess Mortality in the Age of COVID-19: Data Mining for Insights into Death.

1 Motivation

When a global pandemic hits, there are many questions to answer. There is no stopping death wholesale, but how did the COVID-19 Pandemic affect death rates both worldwide and at a more granular scale? How well does excess mortality serve to illustrate the effects of COVID-19? Are any public health policies effective at limiting excess mortality?

2 Literature Survey

A substantial amount of research has been done in regards to the pandemic's effects on mortality rates. Unfortunately, factors such as availability of testing supplies and political influences on public health messaging result in substantial differences between countries and subregions in reporting deaths as attributable to COVID-19 [2, 3, 6]. As a result of these differences public health research has often focused on a different metric: excess mortality, which refers to the number of deaths occurring over and above predicted mortality rates for a population [4]. The variation between reported COVID-19 mortality and excess mortality rates can be staggering: one example country reported 300,000 COVID-19 deaths but had over one million projected deaths over and above expected baseline mortality rates [6]. The majority of prior work focuses on either total excess mortality, a value that is relatively easier to state with confidence, or excess mortality directly attributable to COVID-19.

Predicted mortality rates establish a baseline upon which to compare mortality rates during the pandemic. The four-year period between 2015-2019 is commonly used to create these predicted rates[6, 4, 9], though some models opt to use different periods for a variety of reasons[2, 4, 5]. One of the primary challenges in creating an accurate baseline is the presence of non-pandemic-related factors that affect mortality rates. Changes in population demographics such as age can exhibit a powerful effect on actual mortality[1].

One article in particular, What has happened to non-COVID mortality during the pandemic?[8], does seek to address the same basic question that our project is focused on. While the scope of said article is restricted solely to the United Kingdom, it does provide a possible roadmap for our own inquiries. Another (admittedly pre-print) article presents a possibility for teasing out changes in specific non-

COVID-19 mortality types by projecting current mortality rate subtype units in lieu of local units [7].

3 Proposed Work

3.1 Preprocessing

Both datasets are pre-cleaned with no evidence of missing or erroneous values based on human observation of random samples.

Integration will require work. We need to truncate COVID to include only the 35 countries contained in HMD. Country codes use in each dataset are largely the same, with the exception of Australia, France, Great Britain, Germany, and New Zealand. Additionally, HMD splits Great Britain into three separate categories (England and Wales, Scotland, and Northern Ireland) which will need to be internally merged in HMD before combining with COVID.

Dates between COVID and HMD must also be reconciled. HMD contains weekly data in a YYYY/WW format while COVID contains daily data in a YYYY/MM/DD format. Using pandas, we will resample the daily COVID data to weekly to match HMD (summing "new" data and using the first value for "total" data). At the same time, HMD dates will be converted to YYYY/MM/DD format to aid in searching for specific dates.

Finally, HMD splits mortality data into male, female, and combined categories. To match COVID, the male and female subcategories can be dropped from the dataset.

3.2 Derived Data

Once the datasets are merged, we plan to transform the mortality totals into a smoothed kernel density estimate. This will form the basis of a regression that will inform the expected mortality totals absent COVID after 2020. These expected totals can be compared to reported totals to derive excess mortality.

3.3 Design and Evaluation

We will compare our excess mortality numbers to those of prior work in the field. In addition, we will use expected mortality rates for specific causes (e.g., suicide, traffic fatalities, heart disease) and COVID case numbers to parse a number of questions, including whether published COVID death totals are under- or overcounted and how the rates of other sources of mortality were affected by the pandemic.

4 Data Set

We will be utilizing 3 data sets:

Human Mortality Database (HMD), which contains mortality data by week and age group for 38 countries Accessed at: https://www.mortality.org/, COVID-19 Dataset (COVID) which contains Global COVID Statistics by country and date Accessed at: https://ourworldindata.org/explorers/coronavirus-data-explorer , and COVID-19 Stringency Index which contains time series information on countries and the stringency of their policies regarding the pandemic Accessed at: https://ourworldindata.org/covid-stringency-index

Human Mortality Database

Name	Type	Description
CountryCode	Nominal String	Identifies country
Year	Ordinal Integer	Year in 1 year increments
Week	Ordinal Integer	Week from 1-52 from 1st full week of year
Sex	Nominal String	male, female, and combined
D0_14	Ordinal float	# of deaths age 0 to 14 (float due to transforming input data)
D15_64	Ordinal float	# of deaths age 15 to 64 (float due to transforming input data)
D65_74	Ordinal float	# of deaths age 65 to 74 (float due to transforming input data)

D75_84	Ordinal float	# of deaths age 75 to 84 (float due to transforming input data)
D85p	Ordinal float	# of deaths age 85+
DTotal	Ordinal integer	Total # of deaths for all age groups

COVID-19 Dataset

Name	Type	Description
iso_code	Nominal String	Identifies country
Date	Ordinal String	Date in MM/DD/YYYY format
total_cases	Ordinal Integer	Daily total of COVID cases; does not decrease with recovery or death
new_cases	Ordinal Integer	Daily new cases reported
total_deaths	Ordinal Integer	Daily total of COVID cases
new_deaths	Ordinal Integer	Daily new deaths reported
total_cases_per_million	Ordinal float	total_cases per million
new_cases_per_million	Ordinal float	New_cases per million
hops_patients	Ordinal Integer	Daily total of COVID cases hospitalized; does decrease with recovery or death
population	Ordinal integer	Population of country; does not change

COVID-19 Stringency Index

Name	Type	Description
Entity	Nominal String	Identifies country full name
Code	Nominal String	Identifies country code
Day	Ordinal String	Date in MM/DD/YYYY format
stringency_index	Ordinal Float	The index on any given day is calculated as the mean score of the nine metrics, 0 -100, with 100 being the strictest

5 Evaluation Methods

For now, we think we will be running various regression tests on different categories to see which attributes might affect each other. Keeping in mind that correlation does not imply causation, of course, finding the patterns will lead to more interesting questions that we can explore later. We will also likely be using some sort of clustering technique to see what situations can be compared and contrasted. We might do K-nearest neighbors, although we have not covered that yet in the course material (there’s some linear algebra background here that we can leverage as a team), and we also might treat precaution data as transactional data to see what precautions show up together frequently.

6 Tools

We will be using Python/Pandas to process the data and do most of the heavy lifting. As a team, we are most familiar with that language and have had some prior experience working with Python and Pandas to deal with large data sets. Many things are built in with Pandas, which should make mining for interesting information easier. For visualization's sake, we are interested in working with Tableau. We have seen how powerful that tool can be in the Information Visualization course, and while we are not exactly sure what we want to visualize, we know that Tableau gives us many options. Of course, we will be using GitHub to keep track of changes collaboratively.

7 Milestones

Milestones DONE

14 March - Part 2 due - DONE
16 March - Sample evaluation of US data, overall excess mortality compared to COVID - DONE
23 March - Self Care (Spring Break) - DONE
30 March - Expand evaluation to all countries available in HMD - DONE
06 April - Timeline of COVID public health measures in other countries in HMD - DONE
18 April - Part 3 due (6 pages, progress so far) – DONE

Milestones TO DO

20 April - Expand evaluation of all countries to correlate to COVID public health measures
22 April - Source information on other mortality rates (e.g. suicide, heart disease, traffic accidents)
22 April - Expand evaluation of US data to explore questions of non-COVID mortality
24 April – Finish Final Project Write Up
26 April – Finish Presentation
27 April – Finalize README and submit project code
28 April – Submit Peer Evaluation
29 April – Celebrate!

8 Results so Far

Our datasets have been successfully preprocessed and merged. This was accomplished using the pandas library in python. Both the COVID and HMD datasets were imported as .csv files. The following preprocessing work was performed:

COVID

- Retained only the following attributes: country, date, total_cases, new_cases, total_deaths, new_deaths, total_cases_per_million, hosp_patients, and population
- Dataset was resampled from daily to weekly data. "New" columns were summed and "Total" columns were retained as the first value in each week.
- "year" and "week" columns were extracted from the date column

HMD

- The individual "male" and "female" categories were dropped while the combined category was retained
- Retained only the following attributes: country, week, year, DTotal
- Country codes that differed from the COVID dataset were transformed to match
- Great Britain data, split into England and Wales, Scotland, and Northern Ireland categories, was combined, summing the DTotal
- "week" and "year" were transformed into a "date" column to assist in merging and visualizations

Once both datasets were preprocessed and cleaned, we merged them into a single dataset. We then derived some additional attributes:

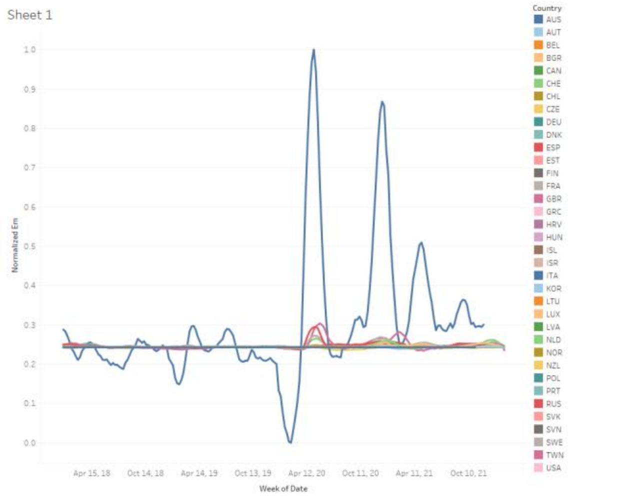
For both weekly deaths and weekly new cases, a moving average was computed by taking the mean of the nearest seven weeks in order to smooth the data and reduce the effects of random variation. Using the smoothed weekly death average for each individual week (e.g., week 1 of all years, week 2 of all years, etc.) we made a linear regression model based on ordinary least squares for each country to predict the number of expected deaths for each combination of week number and year. This was performed by week number to account for cyclical mortality rates that tend to rise and fall based on time of the year.

Subtracting the predicted weekly deaths from the actual deaths gave us an excess mortality number for each country/date combination.

To normalize for population and better compare between countries, we divided excess mortality by population to create the em_per_capita attribute. We further normalized this measure by performing a min-max normalization across all countries.

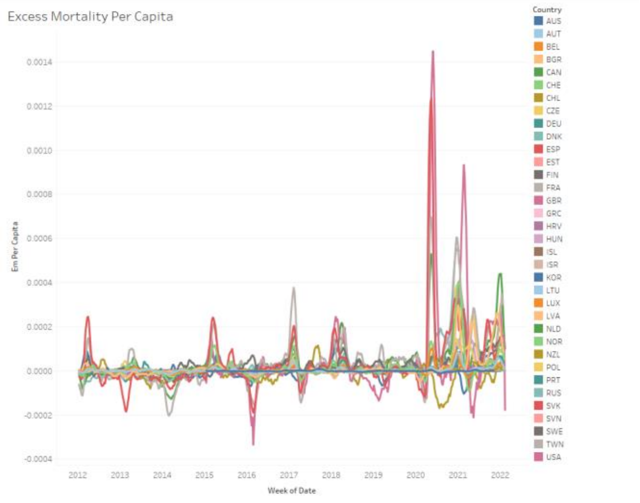
Stringency Index

Because we only have stringency index data from the beginning of the pandemic, another dataframe was made from the previously merged and cleaned datasets. This was done by filtering the Stringency Index data set first by the countries that we have excess mortality data on by excluding country codes that are not present in the cleaned, merged dataset. Then the day was converted to a datetime datatype, in order to pull the week number and year. The indices were then grouped by country code, week number, and year and averaged.

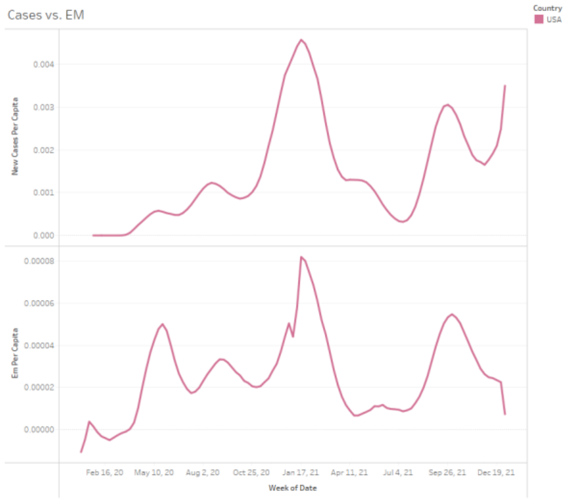


Visualizing the normalized excess mortality rates over time by country, it became clear that Italy’s mortality data was such an outlier that many other countries’ normalized data became indistinguishable between pre-COVID and post-COVID values. For the most part em_per_capita can stand in for the normalized values. For cases where normalized values will be useful, we created a normalization while dropping all Italian rows.

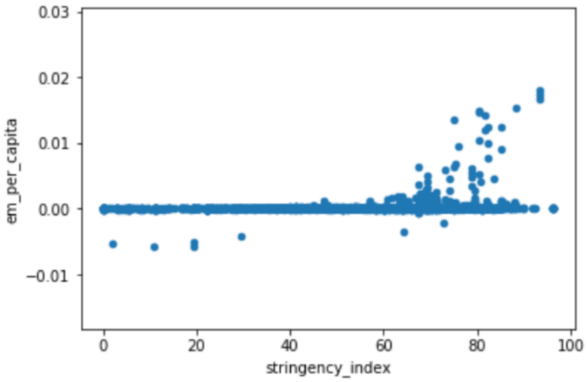
Initial exploratory data analysis using visualizations showed clear evidence of excess mortality during COVID, recreating portions of the past work in this area.



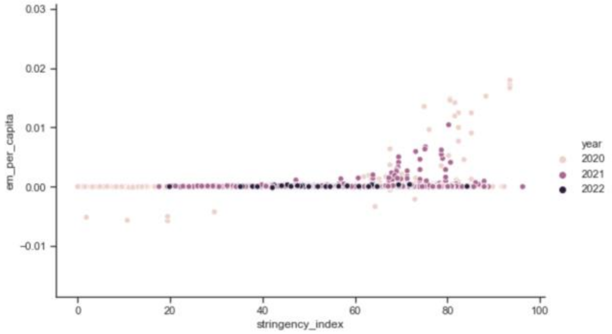
We can also see that excess mortality rates and COVID case rates appear to be correlated.


Cases vs. EM

As for how government regulations affected excess mortality, from this scatterplot, we can see a slight correlation, with a large grouping past 60 on the stringency index.

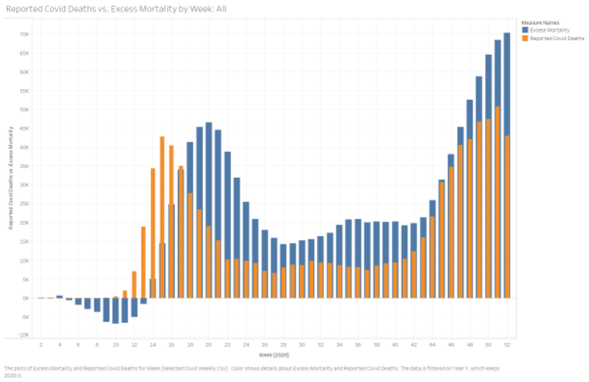


This leads us to examine the dimensionality of the set in a more meaningful way.



By coloring the points by year, we see the farthest reaching happened in year 1 of the pandemic. More filtering, processing, and analysis is needed.

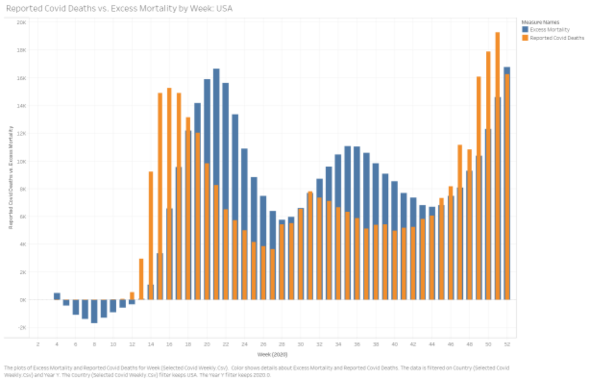
We also have explored visualizations on reported COVID-19 Deaths and Excess Mortality for all countries.


Reported Covid Deaths vs. Excess Mortality by Week: All

The plots of Excess Mortality and Reported Covid Deaths for Week (Selected Covid Weekly.Csv). Color shows details about Excess Mortality and Reported Covid Deaths. The data is filtered on Year Y, which keeps 2020.0.

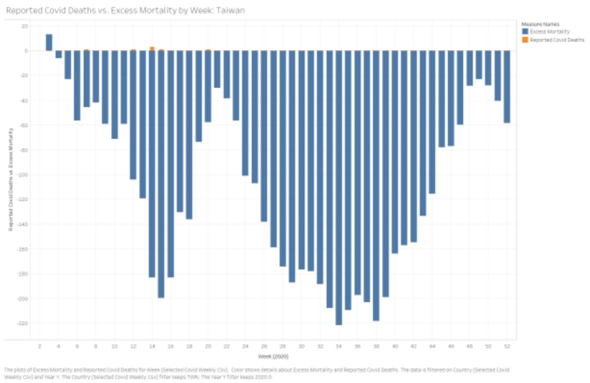
We also have visualizations for these attributes filtered by various countries:

United States:

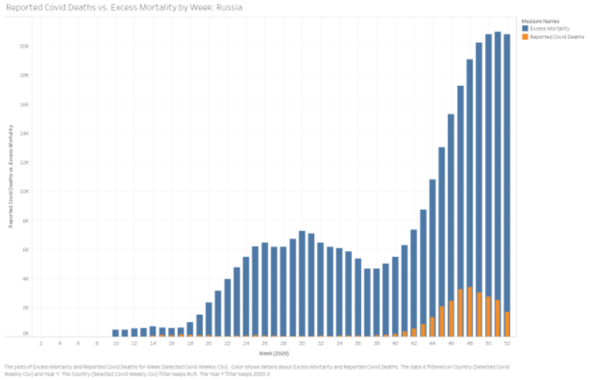

Reported Covid Deaths vs. Excess Mortality by Week: USA

The plots of Excess Mortality and Reported Covid Deaths for Week (Selected Covid Weekly.Csv). Color shows details about Excess Mortality and Reported Covid Deaths. The data is filtered on Country (Selected Covid Weekly.Csv) and Year Y. The Country (Selected Covid Weekly.Csv) filter keeps USA. The Year Y filter keeps 2020.0.

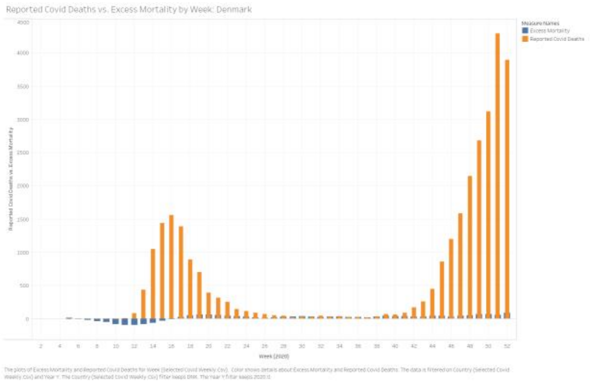
Taiwan:



Russia:



Denmark:



Clearly from the visualizations, there are many different scenarios that need further exploration into the explanation about why the excess mortality and reported COVID-19 deaths are or are not correlated.

REFERENCES

[1] David Adam. 2021. The Effort to Count the Pandemic's Death Toll. In *Nature 601, 312-315 (2022).* https://doi.org/10.1038/d41586-022-00104-8

[2] Abhishek Anand, Justin Sandefur, and Arvind Subramanian. 2021. Three New Estimates of India's All-Cause Excess Mortality during the COVID-19 Pandemic. In *Center for Global Development, Working Paper 589, July 2021, Washington, DC.* https://www.cgdev.org/publication/three-new-estimates-indias-all-cause-excess-mortality-during-covid-19-pandemic

[3] Héctor Pifarré i Arolas, Enrique Acosta, Guillem López-Casasnovas, Adeline Lo, Catia Nicodemo, Tim Riffe, and Mikko Myrskylä. 2021. Years of life lost to COVID-19 in 81 countries. In *Scientific Reports 11, 3504 (2021).* https://doi.org/10.1038/s41598-021-83040-3

[4] Thomas Beaney, Jonathan M Clarke, Vageesh Jain, Amelia Kataria Golestaneh, Gemma Lyons, David Salman, and Azeem Majeed. 2020. Excess mortality: the gold standard in measuring the impact of COVID-19 worldwide? In *Journal of the Royal Society of Medicine. 2020;113(9):329-334*. https://doi.org/10.1177/0141076820956802

[5] Nazrul Islam, Vladimir M Shkolnikov, Rolando J Acosta, Ilya Klimkin, Ichiro Kawachi, Rafael A Irizarry, Gianfranco Alicandro, Kamlesh Khunti, Tom Yates, Dmitri A Jdanov, Martin White, Sarah Lewington, and Ben Lacey. 2021. Excess deaths associated with covid-19 pandemic in 2020: age and sex disaggregated time series analysis in 29 high income countries. In *BMJ 2021;373:n1137*. https://www.bmj.com/content/373/bmj.n1137

[6] Ariel Karlinsky and Dmitry Kobak. 2021. Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. In *eLife 2021;10:e69336*. https://doi.org/10.7554/eLife.69336

[7] Ariel Karlinsky. 2021. National Excess Mortality from Sub-National data: Method and Application for Argentina. Preprint in *medRxiv*. https://doi.org/10.1101/2021.08.30.21262814

[8] Holly Krelle and Charles Tallack. 2021. *What has happened to non-COVID mortality during the pandemic?* The Health Foundation. https://www.health.org.uk/publications/long-reads/what-has-happened-to-non-covid-mortality-during-the-pandemic

[9] Francesco Sanmarchi, Davide Golinelli, Jacopo Lenzi, Francesco Esposito, Angelo Capodici, Chiara Reno, and Dino Gibertoni. Exploring the Gap Between Excess Mortality and COVID-19 Deaths in 67 Countries. In

JAMA Network Open. 2021;4(7):e2117359. https://doi.org/10.1001/jamanetworkopen.2021.17359