**COVID-19: Finding My Own Answers to Avoid Media Propaganda**

Jolene M. Branch

Data Science Program, Bellevue University

DSC 680: Applied Data Science

Prof. Catie Williams

June 27, 2021

**Abstract**

Early in the pandemic, before many Americans even knew anyone who had contracted COVID-19, there was a constant barrage of speculatory ‘news’ about the expected trajectory of the virus, anticipated effects of cold weather upon viability of the virus, assurances that masks were not necessary to prevent transmission among asymptomatic members of the general public, and dire warnings from citizens in Italy that Americans should prepare lest the illness and deaths in their country be repeated on an even larger scale in the United States. Some of these predictions were attributed to historical accounts of epidemiology of other viruses while others were suspected immediately by members of the medical community to have ulterior motives (such as announcements that only healthcare workers needed to wear masks and that 40,000 ventilators were needed immediately in New York City). This paper attempts to answer five research questions; “Does COVID-19 ‘survive the winter’ and have higher prevalence rates in countries with higher latitudes?”, “Are COVID-19 death rates lower in countries that do better at preventing certain chronic conditions?”, “Do countries with higher population densities have higher COVID-19 prevalence?”, “Is there a correlation between case counts of SARS/MERS and COVID-19 prevalence rates in countries?” and “Do countries with more people who have had COVID (prevalence) have higher death rates (like, maybe due to additional strain on healthcare system)?” Data that has been collected by the most reliable sources possible and handled the least will be used, requiring extensive data ‘wrangling’ in R prior to exploratory data analysis, insight generation, and statistical modeling. The bulk of the coding for this project is done within the data clean-up part of the process, using R.

*Keywords:* COVID-19, healthcare, epidemiology, fake news, incidence, prevalence

**COVID-19: Finding My Own Answers to Avoid Media Propaganda**

**Purpose**

The purpose of this project is to retrieve and wrangle the data myself in order to come to my own conclusions about what the COVID-19 virus is doing. A source of unbiased reporting of COVID-19 information is not known. However, unbiased *data* is available.

**Data Sources**

Three datasets from the Johns Hopkins Whiting School of Engineering Github repository (<https://github.com/CSSEGISandData>) were used. The datasets were in the form of .rds files. One contained COVID case counts and new cases by country. Another contained MERS and SARS data for the same countries, along with environmental and demographic information. A third contained latitude, longitude, and populations. All were imported into R using readRDS().

**Methods**

After importing the datasets into R, they were written back out as .csv files for simpler importation into SQLite. (Microsoft SQL Server Management Studio was not readily available for use). After trial and error of joining datasets within SQLite and the realization that at least one dataset needed to be changed from long to wide format, all three were brought back into R for reshaping and joining (using merge()).

***Variable Selection***

Variables were selected based on ease of retrieval, availability of data definitions, and clinical understanding of chronic illness and the impact of the external factors, such as national healthcare delivery models, population density, as well as predisposing conditions (such as obesity) on chronic conditions. In the event that enough data would be available for predictive analytic modeling, SARS and MERS prevalence data was included, as “…the lessons we learned in the past from the SARS and the MERS epidemics are the best cultural weapons with which to face this new global threat.” (Petrosillo et al., 2020). Interestingly, China, which is known to be the country of origin of the COVID-19 virus, had the highest count of SARS cases.

***Data Cleaning and Dataset Trimming***

Because COVID-19 data has been fastidiously collected on a daily basis by organizations such as Johns Hopkins, many publicly available datasets are of such great size as to make personal handling unfeasible (Kellen & Maragakis, N.D.). This was mitigated by taking a single point in time snapshot of COVID-19 data across the world. The date selected was June 11, 2021. Fortunately, epidemiologic data nearly always contains ‘cumulative’ data and/or prevalence data, so even a single point in time scan of available data reveals much more than what is occurring on just that day.

After merging the three datasets, cleaning and wrangling the data, and spot checking for sanity, the combined dataset was imported into Power BI. Calculated columns of infectious disease prevalence rates of COVID-19, MERS, and SARS were created next. The validity of the MERS and SARS data is questionable due to extremely small case counts, as neither of those diseases made it to the ‘pandemic’ level of spread.

The need to normalize the data became clear when almost no correlation between deaths and COVID-19 cases per 100,000 people was found, as well as a negative correlation between chronic disease and COVID-19 cases per 100,000. Counts needed to be converted to rates per 100,000.

***Outlier Detection and Missing Value Decisions***

The normal process of outlier detection was viewed differently within this project than within most data science projects. Traditionally, each variable in a dataset is examined, often via visual representations such as histograms, cumulative distributive function (CDF), or probability mass frequency (PMF) charts. In the case of a brand new rapidly emerging virus, however, outliers are to be expected and attempts to remove them introduce the risk of destroying essential elements of insight within the dataset. The matter of miniscule sample sizes (from countries with population of less than 100,000, for example) and their propensity to skew data was decided to be too likely to introduce noise into the understanding of COVID-19, and those countries were removed from the datasets prior to analysis. Unfortunately, many of the countries with populations of less than 100,000 share other characteristics, such as island status and location in warmer climates (or nearer to the equator).

Additionally, countries that were not able to supply data on such commonly collected and reported variables such as obesity, smoking, and HIV status were eliminated from the datasets due to concerns about credibility of any of the supplied data given the glaring omissions of ‘standard’ public health data. These attempts to decrease noise within the data likely served to remove demographically similar countries from the dataset, which does decrease applicability (or decrease generalizability) across all countries.

Out of data modeling convenience, the U.S. was compared against only Eastern Northern Hemisphere countries for the final round of correlations and analysis.

**Analysis and Findings**

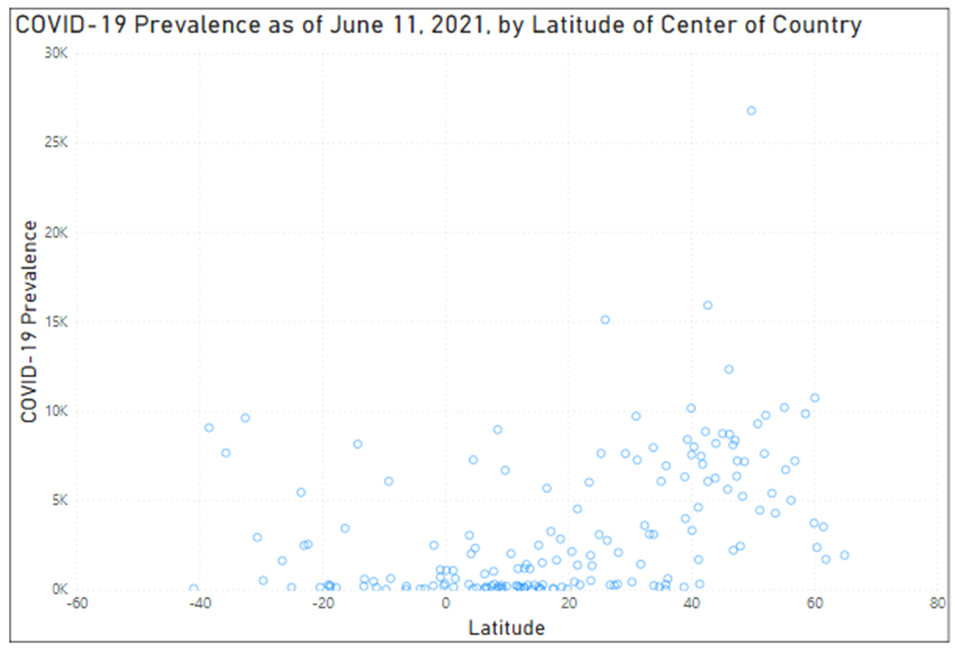
***Question 1: “Does COVID-19 ‘survive the winter’ and have higher prevalence rates in countries with higher latitudes?”***

Early in the COVID-19 pandemic, there was speculation (and hope) that the virus would have different survival rates depending on geographic locale. There was speculation that the virus would not survive as well at higher altitudes or further away from the equator.

Timeline, map

Description automatically generated *Note:* Larger circle size indicates higher prevalence of COVID-19.

Since distance from the equator can be calculated as the absolute value of the latitude, a latitude of 40 is the same distance from the equator as a latitude of -40. From this chart a sort of ‘V’ shape among rates of COVID-19 prevalence as compared to latitude of the center of the country. This suggests a correlation between latitude and COVID-19 prevalence.

 *Note*: Each circle represents one country.

The prevalence of COVID-19 in countries as of June 11, 2021 appears to have a positive correlation with latitude.The countries farther from the equator, in general, had higher COVID-19 cumulative rates. However, for sheer counts, the opposite relationship was seen. The countries farther from the equator had fewer cases. This is likely due to lack of adjustment for population when a straight case count is used, making that relationship not helpful.

***Question 2: “Are COVID-19 death rates lower in countries that do better at preventing certain chronic conditions?”***

Chart, bubble chart

Description automatically generated

*Note*: C19.100k indicates number of confirmed cases per 100,000 people on this correlation matrix.

Negative correlations are in blue and positive ones are in red. Yes COVID-19 death rates are higher in countries that do poorer at prevention of chronic medical conditions. In the correlation matrix, COVID deaths have a positive correlation with every chronic condition listed *except diabetes*. The only variable on the correlation matrix that does not have a positive correlation with death, interestingly, is diabetes. Not even in the media has it been suggested that having diabetes is some sort of protective factor against COVID-19! Since diabetes and obesity do not show as having a positive correlation (which is simply common knowledge), this is a clue that something is amiss with the diabetes variable within this dataset.

***Question 3: “Do countries with higher population densities have higher COVID-19 prevalence?”***

The correlation matrix does not support this statement.

***Question 4: “Is there a correlation between case counts of SARS/MERS and COVID-19 prevalence rates in countries?”***

Upon review of SARS and MERS data as collected by the Johns Hopkins Whiting School of Engineering, it was discovered that so few countries had formally reported actual cases of SARS and MERS, that valid comparison with COVID-19 case prevalence would not be possible. There simply is not enough case count data.

***Question 5: “Do countries with more people who have had COVID (prevalence) have higher death rates (like, maybe due to additional strain on healthcare system)?”***

There is not strong enough correlation within the dataset to support this statement.

**Conclusions**

Initial hypotheses (in the form of questions) proved to be as likely to be correct as incorrect. This is thought to be attributed to the constant barrage of COVID-19 related ‘news,’ and has great potential to lead individuals, local government, and state government to be initiating conversations based on inaccurate data. Lack of trustworthy data is a potential contributing factor to continued spread of COVID-19 globally.

**References**

Asrani, P., Hasan, G. M., Sohal, S. S., & Hassan, M. D. I. (2020, November 24). *Molecular Basis of Pathogenesis of Coronaviruses: A Comparative Genomics Approach to Planetary Health to Prevent Zoonotic Outbreaks in the 21st Century*. Omics : a journal of integrative biology. <https://pubmed.ncbi.nlm.nih.gov/32940573/>

Bernstein, B. (2021, April 17). *COVID-19 Death Toll Surpasses 3 Million Worldwide*. National Review. <https://www.nationalreview.com/news/covid-19-death-toll-surpasses-3-million-worldwide/>

D'Agostino, A., Demartini, B., Cavallotti, S., & Gambini, O. (2020, May). *Mental health services in Italy during the COVID-19 outbreak*. The Lancet: Psychiatry. <https://doi.org/10.1016/S2215-0366(20)30133-4>

Dong, E., Du, H., & Gardner, L. (2020, February 19). *An interactive web-based dashboard to track COVID-19 in real time*. Redirecting. <https://doi.org/10.1016/S1473-3099(20)30120-1>

Elrashdy, F., Redwan, E. M., & Uversky, V. N. (2020, September 11). *Why COVID-19 Transmission Is More Efficient and Aggressive Than Viral Transmission in Previous Coronavirus Epidemics?* Biomolecules. <https://pubmed.ncbi.nlm.nih.gov/32933047/>

El-Sayed, A., Abdel-Daim, M. M., & Kamel, M. (2021, May 5). *Causes of respiratory failure in COVID-19 patients*. Environmental science and pollution research international. <https://pubmed.ncbi.nlm.nih.gov/33954917/>

Kellen, G. D., & Maragakis, L. L. (n.d.). *COVID-19 Vaccine: What You Need to Know*. Johns Hopkins Medicine. <https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/covid-19-vaccine-what-you-need-to-know>

Liu, Z., Xiao, X., Wei, X., Li, J., Tan, H., Zhu, J., Zhang, Q., Wu, J., & Liu, L. (2020, March 11). *Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2*. Journal of medical virology. <https://pubmed.ncbi.nlm.nih.gov/32100877/>

Petrosillo, N., Viceconte, G., Ergonul, O., Ippolito, G., & Petersen, E. (2020, March 28). *COVID-19, SARS and MERS: are they closely related?* Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. <https://pubmed.ncbi.nlm.nih.gov/32234451/>

Singh, R., Kang, A., Luo, X., Jeyanathan, M., Gillgrass, A., Afkhami, S., & Xing, Z. (2021, March). *COVID-19: Current knowledge in clinical features, immunological responses, and vaccine development*. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. <https://pubmed.ncbi.nlm.nih.gov/33577115/>

Yang, Y., Rao, X., & Shang, W. (2020, June). *Facing the COVID-19 outbreak: What should we know and what could we do?* Journal of medical virology. <https://pubmed.ncbi.nlm.nih.gov/32091134/>

**Appendix**

Chart

Description automatically generated

Treemap chart

Description automatically generated