Evaluation of The Impact of COVID19 with a Simple Data Analysis

Vincent Chu - 1003938778

23 January, 2022

Abstract

With the outbreak of COVID-19, the issue of misinformation is brought to the attention of many citizens and policymakers. This report aims to address the piece of misinformation which suggests "COVID is just another flu" with a simple data analysis, so it can be understood by a general audience. The dataset used in this report is the "Provisional weekly death counts, by age group and sex" from Statistics Canada. This dataset includes information on location, time and the weekly death count, and it will be used to compare the trends of death count in provinces with varying levels of impact from COVID-19. The difference in difference approach will be used since the provinces within the Atlantic Bubble serves as a perfect comparison since they have experienced a prolonged period of 0 COVID-19 cases. The model found statistically significant results in the association of COVID-19 and increased death count. The report addressed some major threats to the validity of the difference in difference model. However, causal inference cannot be made as there still exists some confounders which are not accounted for in the model.

Keywords: Difference in Difference, COVID-19, Misinformation, Canada, Death, Mortality

Introduction

The current pandemic caused by the COVID-19 virus has been a pertinent and debated topic for a number of reasons . Many have expressed their concerns for the outlook of the economy as more and more economic activities have been shut down or forced to operate in a reduced capacity to limit the spread of the virus, [6]. Some are concerned about ethical and human rights issues, especially regarding individual freedom due to the tightening restrictions imposed by government policies in hopes of containing the virus outbreak [7]. However, there is one particularly prominent issue that has been brought under the spotlight by COVID-19. Misinformation has become a greater issue over the years as an increasing percentage of consumers started using social media as a source for news, and technologies such as deepfakes begin to mature [8] [9]. This issue has caught the attention of many citizens as well as policy makers as COVID-19 related misinformation impacts the vaccination effort on a global scale [4][5]. This is likely due to the effect that misinformation has on undermining the vaccine effort, which results in a continuous burden on the Canadian healthcare system. This also puts a strain on the provinces' plans to reopen as these plans are heavily dependent upon vaccination rate [13] [14]. According to Loomba et al (2021), there is a statistically significant difference in an individual's intent to accept the COVID-19 vaccine after being exposed to misinformation [12].

The goal of this report is to address one of the popular pieces of misinformation using a straightforward statistical technique so a general audience with minimal statistical background is still able to have a comprehension of the data analysis approach. The piece of misinformation that will be addressed is "COVID-19 is just another common flu" [11].

This piece of misinformation downplayed the seriousness of COVID-19 and thus the importance of the vaccine. It is widely acknowledged that the spread of this piece of misinformation can be attributed to the

former US President Donald Trump's Tweet [10]. Another possible reason for the spread of this misinformation is the confusion about the difference between coronavirus and COVID-19. COVID-19 is a specific strain of coronavirus, much different from the generic coronavirus that causes the common cold [23]. In order to have a better comprehension of the impact of COVID-19, this report will utilize the "Provisional weekly death counts, by age group and sex" dataset from StatCan[1]. Using this dataset, an analysis on the trends of the provincial weekly death count will be performed using a method called "difference in difference". This method can be used to compare the change in pattern of death count in Eastern Provinces and other provinces in order to answer the question of whether COVID-19 is just another common flu. It is hypothesized that the trends of death counts of all provinces were similar before COVID-19 given the same age category, since the Canada Health Act ensures the standard of access to healthcare across Canada [15]. Additionally, it is hypothesized that the trends of death count between non-Eastern Provinces and the Eastern Provinces will begin to differ with the introduction of COVID-19 with Eastern Provinces having a lower increase in the trend of death count when compared to non-Eastern Provinces.

In this report, Eastern Provinces refers to the provinces within the Atlantic Bubble, which are Nova Scotia, Prince Edward Island, New Brunswick, and Newfoundland and Labrador [31]. The Eastern Provinces are used as a comparison to other provinces since they are a part of Canada that has recorded a prolonged period of 0 COVID-19 cases. They also share many similarities in history, culture, healthcare standards, and etc, with the rest of the Canadian provinces. It would be logical to assume that these provinces are similar in nature as well as their weekly death count trend when compared to the other Canadian provinces. If the trends of death counts between the provinces are identical before COVID-19, which is also known as the parallel trend assumption, the Eastern Provinces would be ideal to be used to estimate the counterfactual of non-Eastern Provinces death count trends without COVID-19.

Data

Data Collection Process

As mentioned in the introduction, the dataset that will be used in this report is called " Provisional weekly death counts, by age group and sex"[1]. This dataset is published by Statistics Canada. The goal of this dataset is to create a dataset that can be used to examine the deaths within Canada in relation to the quality of healthcare in the location of their death. Statistics Canada collects death data of Canandians using the Canadian Vital Statistics Death Database (CVSD). Deaths are reported on a provincial or territorial level using their respective Vital Statistics Death Database; the data is then reported to Statistics Canada, forming the CVSD [28]. This system provides data to researchers, epidemiologists, clinicians, and policymakers to respond to urgent public health matters, for example, flu, suicide, and the opiod overdose crisis[28].

This dataset is acquired from StatCan using the statcanR package in R [36]. The dataset contains 127890 observations and 18 variables. It contains information on the weekly death count within Canada starting from 2010. It also contains the demographics of the deceased, like location, age, and sex.

Cleaning Process

The dataset contains 18 variables, many of which irrelevant or empty columns. Therefore, the first step in the cleaning process is to remove the irrelevant variables by creating a subset that contains only the following variables: dates, location, age at time of death, sex, value. The description of the variables within the subset is as follows:

- Dates: the date of the last day of the specific week, which is a Saturday.
- Location: Categorical variables which describe the location by the death count value in terms of provinces, territories or whether the death count is on a national scale. There are 14 values for this variable, which consist of the 10 provinces, 3 territories and 1 value for Canada.

- Age: Categorical variables that describe the age of the deceased at the time of death, consisting of 5 categories. Age 0 to 44, age 45 to 64, age 65 to 84, age 85 and over, as well as all ages.
- Sex: Categorical variable consisiting of three categories male, female, both sexes
- Value: Discrete variable that describes weekly death counts for the given demography.

The difference in difference model aims to compare trends of weekly death count in different locations. The dataset is filtered to only contain observations which the sex variable contains the value of "both sexes" since the subject of interest is death count regardless of sex.

Table 1: Summary Statistics

Province	**Average Weekly Death Count**	**Missing Value**
Alberta	189.	0
British Columbia	277.	0
Canada	2053.	15
Manitoba	82.2	175
New Brunswick	54.7	41
Newfoundland and Labrador	38.4	0
Northwest Territories	1.57	368
Nova Scotia	70.4	25
Nunavut	1.00	379
Ontario	766.	25
Prince Edward Island	9.75	58
Quebec	496.	25
Saskatchewan	73.3	15
Yukon	1.51	1,240

Table 1 summarizes the weekly death count by location as well as age category. Since the weekly death count is a raw number but not a ratio to the population of the province, the provinces with a larger population would naturally have a higher value. This would affect the difference in difference model as the same percentage change in Ontario and Alberta would result in a drastically different slope. To account for this difference, the weekly death count value will be normalized according to its historical mean. The normalized weekly death count would reflect on the percentage deviation from the mean. Another observation is that Nunavut and Northwest Territories have vast amounts of empty data. In general because of the low population, the territories have extremely low weekly death counts for the younger age category. This might be a concern if there are many zeros in the data as sparse data might negatively affect the data's ability to form trends.

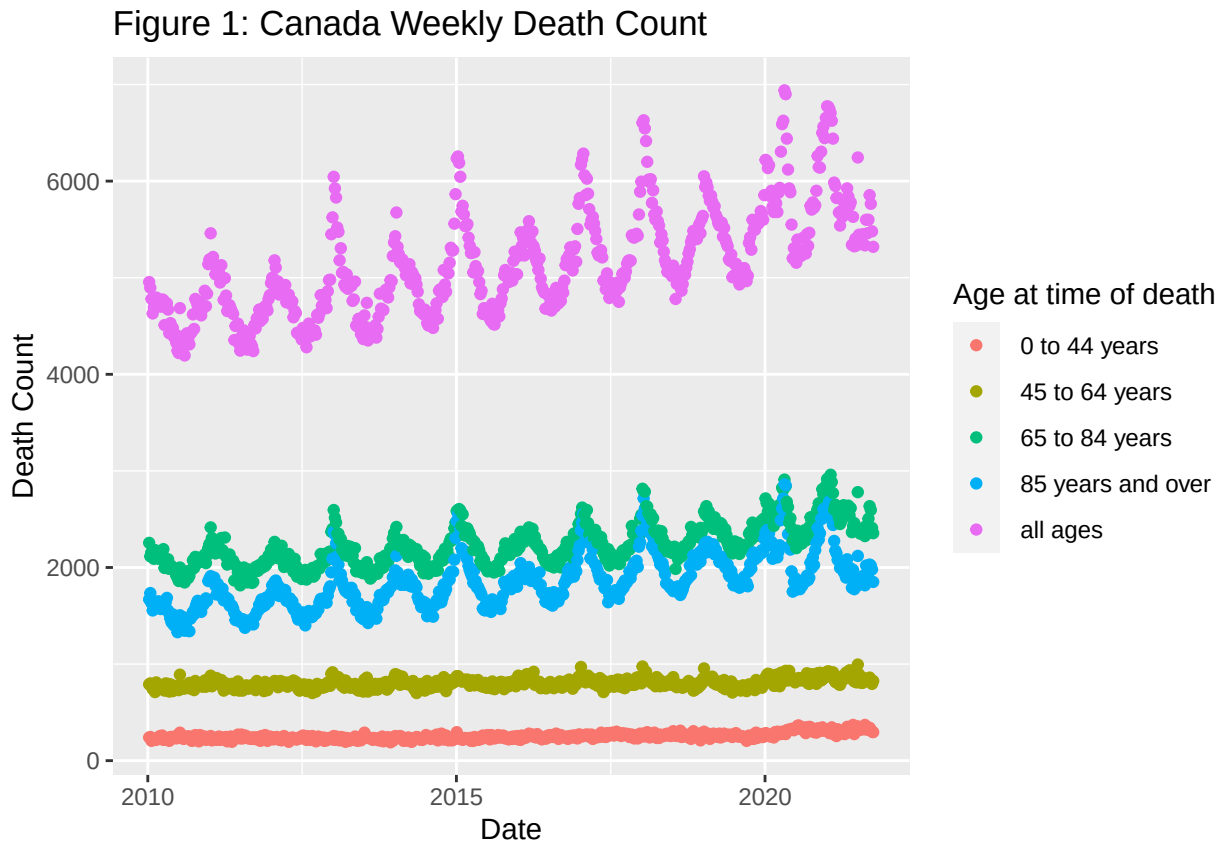
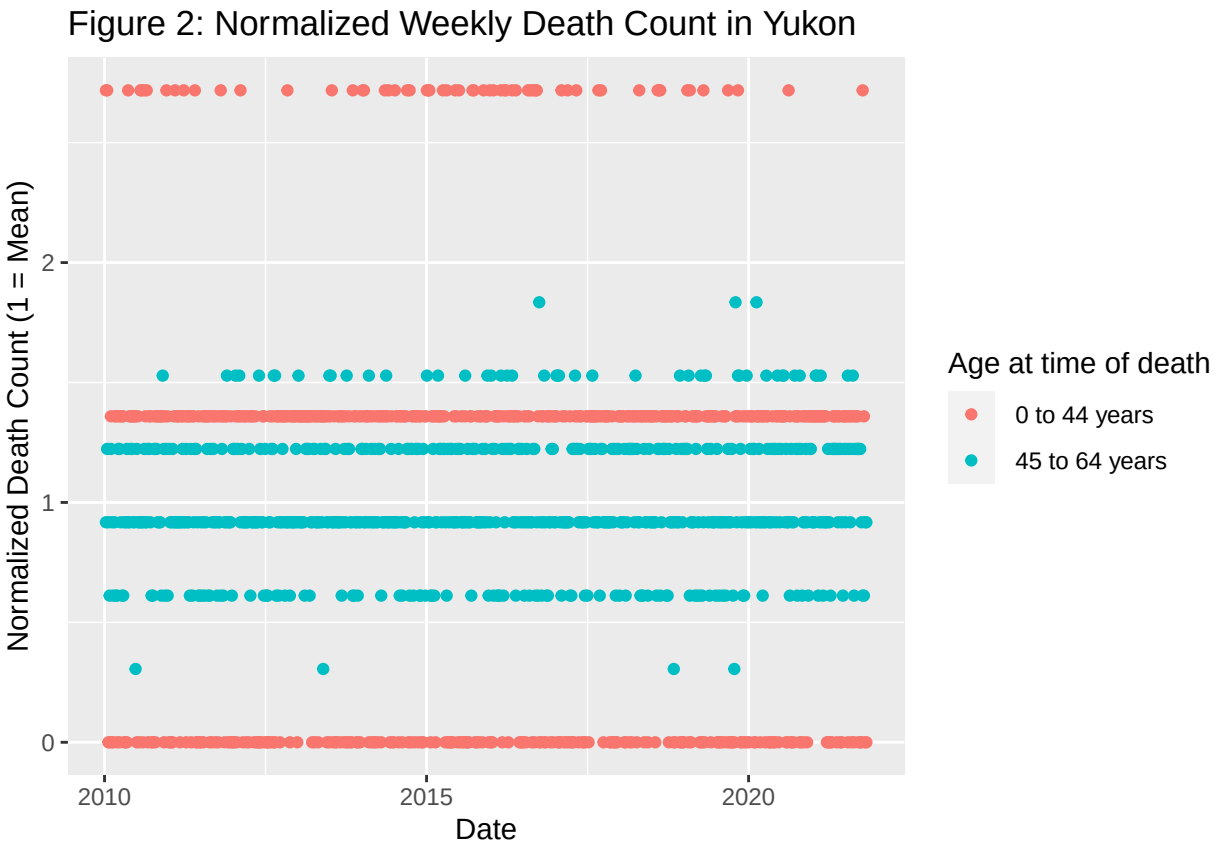


Figure 1 features a visualization of the weekly death count in Canada by age group. This graph is particularly interesting because of its unique pattern. The overall death count in Ontario is increasing despite healthcare technologies and living standard improvements. This is likely due to the aging and growing population of the country [17]. There is also a noticeable spike of death count every year around the December and January mark. This is likely due to the seasonal flu [20]. Interestingly, the "65 to 84" and "85 and above" age groups are almost the sole contributor to the spike in death count. This also means the weekly death count of those two age groups do not exhibit a clear linear relationship with respect to time. This would make it more difficult to examine the parallel trend assumptions. With the information from Figure 1, it would be wise to compare the weekly death count trend of one of the two younger age groups between different provinces, since the weekly death count value of these two categories exhibit a more linear trend. Hence the data will be further reduced to a subset containing only the two younger age categories.

The next step will be to normalize the death count by province. This step is necessary because of the nature of the dataset, the weekly death count is the raw value and is not a ratio of the provincial population, as shown in Table 1. For example, a one-percent increase in the value of weekly death count in Ontario results in a much larger increase in the coefficient of the slope of the trend when compared to a less populated province. To account for this difference, the weekly death count value will be normalized according to its historical mean. Normalizing the death count will adjust for the difference in the total population of the provinces and effectively display the changes as a percentage from the mean, which is needed to investigate variables of interest, the change in the weekly death count.



From Figure 2, a clear issue can be seen after normalizing the death count value. The Eastern Provinces as well as the territories have an extremely small population. This led to the weekly death count normalized value for the youngest category often ending up being zero. Using Yukon as an example, the weekly death count has 1,240non-empty observations of which 334 of the age category “0 to 44” are zeros and 290 of the age category “45 to 64” are zeros. The data for Northwest Territories exhibits a similar behaviour. This is problematic because the mean of the death count value would be extremely low and the normalized weekly death count will remain at zero for the majority of the time and consist of an extremely high value when the death count is not zero. In addition, Nunavut stopped updating their weekly death count after 2016. Therefore, all territories will be removed. In terms of the eastern provinces, Prince Edward Island is the only one that exhibits this issue, and therefore it will be removed from the dataset as well.

Figure 2 demonstrated the effect of sparse data when performing normalization and the overall impact on the ability to analyze the trend itself. As seen from Figure 1, the “45 to 64 years” age category has a higher weekly death count when compared to the “0 to 44 years” age category. Figure 3 aims to examine whether the ”0 to 44 years” age category should be included in the difference in difference model.

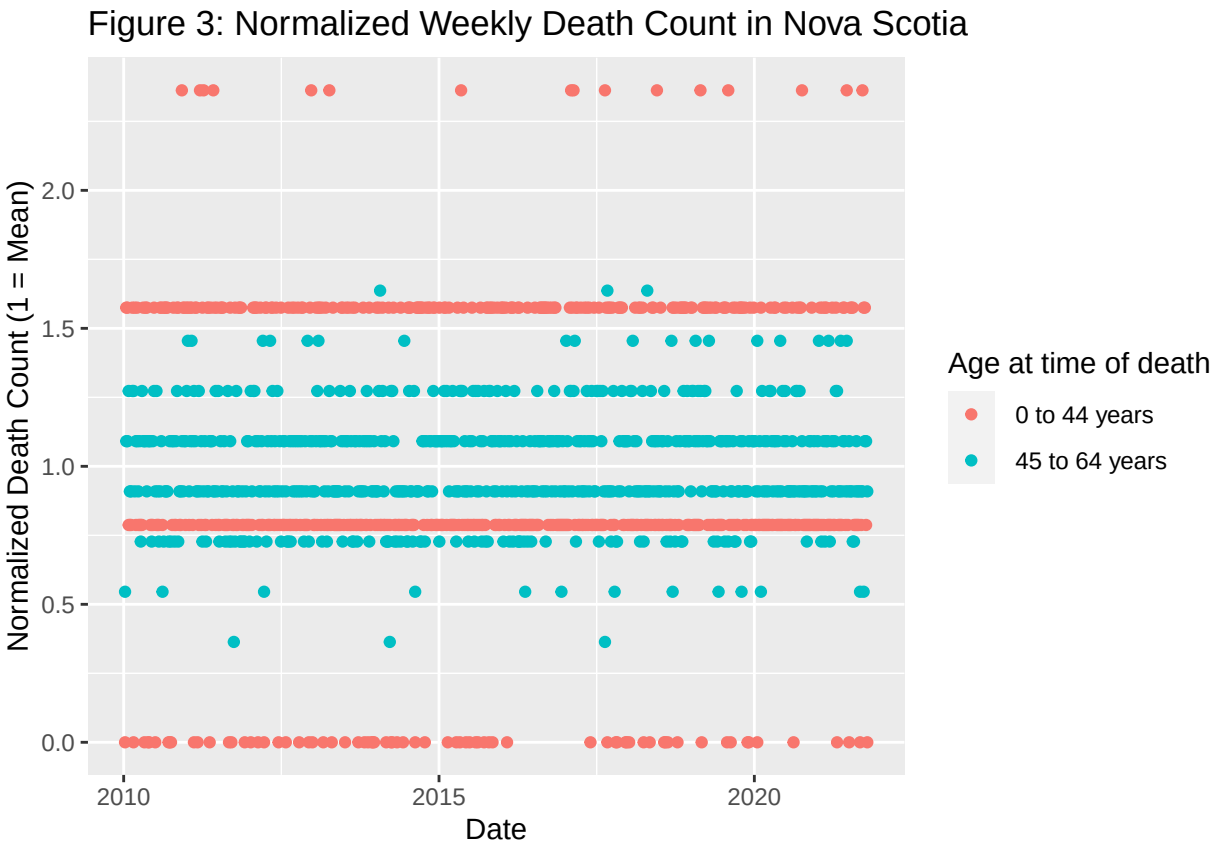


Figure 3 displayed the normalized weekly death count for Nova Scotia, which is the largest eastern province by population. It clearly shows that the "0 to 44 years" age category exhibits the issue of a high amount of zeros and there is little variation of the weekly death count value. This suggests that the weekly death count of "0 to 44 years" age category might not be suitable to be used for the difference in difference model.

The final subset of the data consists of date, value, and location. The variable sex and age has been removed since there is only one category in each of the variable. Sex being both sexes, and age being "45 to 64 years". In terms of location, the territories and Prince Edward Island have been removed due to sparse data that makes it difficult to analyze trends especially after normalization.

The last step in the data cleaning process will be to add the dummy variables. There are two dummies that will be added, treatment and post. The treatment dummy represents whether a province is affected by COVID-19. Only the Eastern Provinces which are within the Atlantic Bubble will be assigned the value 0, as they are the only provinces in Canada to have a prolonged period of 0 cases and have only experienced very minor outbreaks in the recent month[32]. All the other provinces will be assigned value 1, as they are affected by COVID-19. The post dummy represents whether the period of time is before and after COVID-19. There is speculation that COVID-19 has existed in Canada before testing for the virus began as the seriousness of the virus was underestimated [22]. However, setting the date arbitrarily is susceptible to p-hacking and therefore the cutoff date will be set on January 27, 2020 [21]. This is the date of which the National Microbiology Lab confirms a person undergoing quarantine at Sunnybrook Hospital is the first documented confirmed case of COVID-19 in Canada [21]. The end of the date range is also narrowed to January 2021 as the vaccination program for COVID-19 begins in Canada [30]. As the vaccine can effectively reduce the lethality of COVID-19 as well as reducing its spread, the impact of covid will be undermined [29].

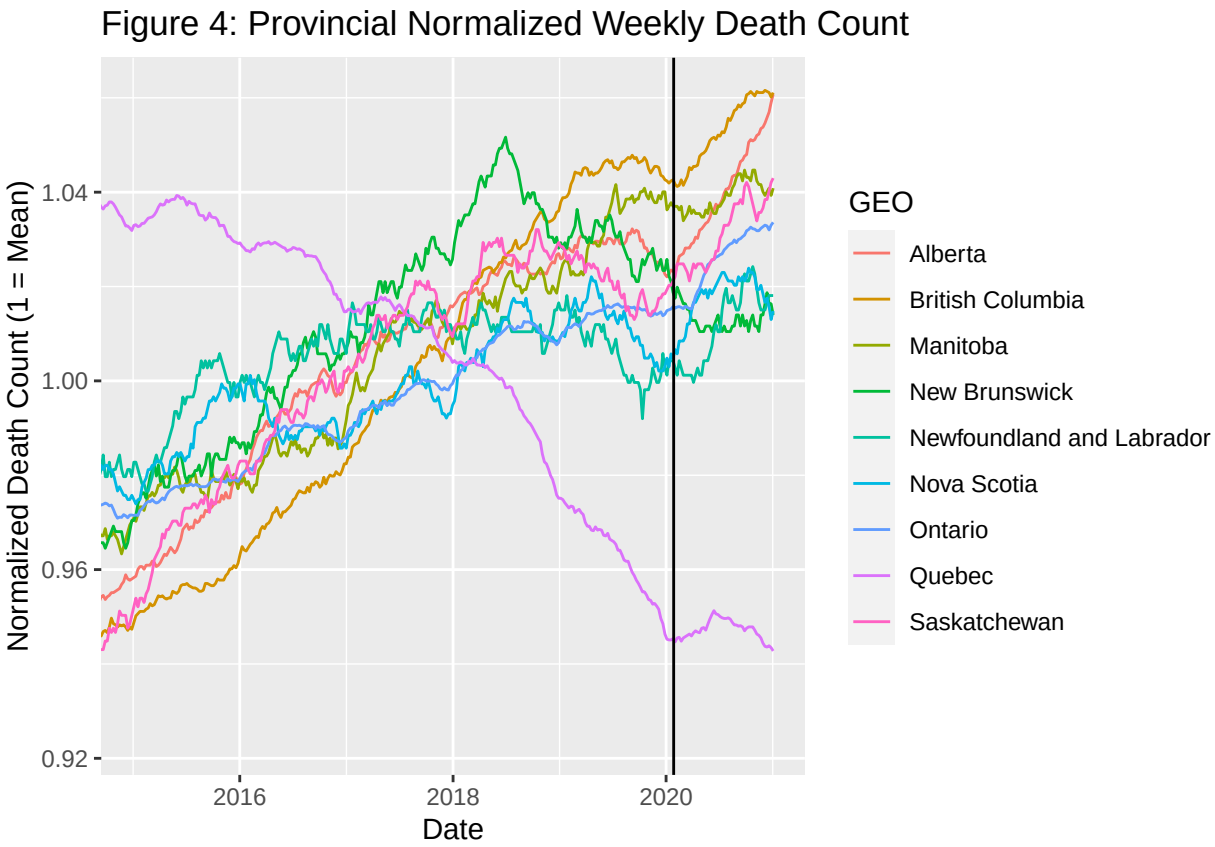


Figure 4 shows the 200 week moving average of normalized weekly death count by province for the age category "45 to 64 years" from January 2015 to December 2020. This graph visualizes the trend of weekly death count in different provinces and it is helpful in terms of evaluating the parallel trend assumption. There are 5166 data points in the scatterplot. To make the table readable, a slightly smoothed line is used instead for the visualization of the data. According to this graph, the value of the weekly death count has been on the rise since 2011 for all provinces shown in this graph with the exception of Quebec. The trends between the Eastern Provinces and non-Eastern Provinces move together in a very similar manner with the exception of Quebec once again. With the information provided by this graph, Quebec should be removed from the dataset since the death count trend of the provinces does not follow the parallel trend assumption. This is reasonable since Quebec is different from the rest of Canada because they have a distinct French culture that the rest of Canada does not share. Quebec law is also unique in Canada as it is the only province in Canada with a civil code [26]. Other than Quebec, the parallel trend assumption for death count data are similar across other provinces shown in the Figure and the parallel trend assumption can be considered satisfied.

Summary of Variables

The original dataset has been inspected and cleaned. At first, irrelevant variables are filtered and only date, location, value, sex, and age remains. The varaible sex is then filterd to only have observations belonging to "both sexes" as the interest of this report is the overall death count. Upon the initial visualization of the weekly death value on a national scale, it is clear that the weekly death count of the two older age categories does not exhibit a linear relationship with time. The death count value seems to be seasonal and consist of a yearly spike around December and January. It would be very difficult to evaluate the common trend assumption for these two age categories. Therefore, the two older age categories are filtered out. The death count value is then normalized to account for the difference in population between the provinces. The weekly death count value for the two youngest age categories contains a large number of zeros for provinces and

territories with lower populations. Due to the large number of zeros, the data does not exhibit a clear trend and has a huge fluctuation in value after normalization,.and because of this reason the three territories and PEI are removed. However, the remaining provinces still suffer from the issues with large numbers of 0s in the data for the youngest age category. The youngest age category is removed as well, leaving the second youngest age category "45 to 64 years" as the only age category. And lastly, three dummy variables are added. The treated groups dummy assign a value of 1 to those provinces which are affected by COVID-19, which are the provinces outside of the Atlantic Bubble. The post dummy describe whether COVID has been introduced in Canada, the date is set to January 27, 2020 as it is the first day of a confirmed case. The last dummy variable is treated which describes whether the province was affected by COVID-19 at a specific time. The end of the date variable has also been limited to December 2020 as vaccines are being introduced in Canada, which would undermine the effect of COVID-19.

The final dataframe describes the weekly normalized death value for the demographic with age "45 to 64 years" and both sexes. The variable date is a discrete variable that describes the date of the last day of a specific week, from January 9, 2010 to January 2, 2021. Location is a categorical variable with 9 different categories representing 8 provinces, excluding the territories, Prince Edward Island, and Quebec. Death is a continuous variable that describes the value of the weekly death count normalized by historic mean. Treated group is a dummy variable for whether the province is affected byCOVID-19. Post is a variable for whether the data exists after the introduction of COVID-19 to canada. The product of treated group and post is the variable treated which represents whether the province has experienced COVID-19. Note: variable age and sex are not included in the final dataset as they only consist of one value.

Methods

Difference in differences is a statistical technique that makes use of observational data with the goal of forming causal inference. Difference in difference works by comparing the change in outcome of the control and treatment group over time. This is done by using the control group to approximate the counterfactual of the treatment group. Since difference in difference is effective in removing unobserved confounders, it is often used to form causal inference in data when randomization on an individual level is not possible. In the context of COVID-19 and death, it is almost certain that executing a randomized experiment is unethical. Given the structure of the dataset being time series data with appropriate groups that can be categorized as treatment and control. Treatment group being provinces outside of the Atlantic Bubble, and the control group being the provinces within the Atlantic Bubble. The use of difference in difference would be an appropriate model to determine the impact of COVID-19 on death trends in Canada given the assumption of the model being satisfied.

The "Provisional weekly death counts, by age group and sex" dataset used in this report is an observational dataset. This dataset has the weekly death count data for each province, seperated by age and sex. Difference in difference would be an appropriate approach since the Eastern Provinces in Canada had a prolonged period of time of zero cases. Even in the case of an outbreak, the case numbers are minor and the Eastern Provinces are able to quickly put outbreaks under control [18]. The Eastern Provinces will be ideal to act as the candidates for the control group, since they are the part of Canada that experience little to no cases. A difference in difference model with the eastern provinces as control group and the rest of the provinces as treatment group is very likely to satisfy the parallel trend assumption.

Assumptions

The parallel trend assumption states that the difference between the treatment and control group is constant over time. This assumption is crucial as it allows the calculation of the trend that would have occurred to the treatment group if the treatment group did not receive treatment, which is also known as the counterfactual.

This assumption can be examined in the visualization depicted in Figure 4. In Figure 4, It is shown that all the provinces displayed in the figure, with the exception of Quebec, have similar changes in weekly death

count. Please note that the trends displayed in Figure 4 are not completely identical across provinces, but the difference is extremely minor with a difference of less than 1% of the historic mean. Therefore, the parallel trend assumption is considered to be satisfied by the provinces displayed, with the exception of Quebec.

The selection of a data is solely dependent on whether the province satisfies the parallel trend assumption. In the case of age categories, the two older age categories do not exhibit a linear trend with respect to time as the weekly death count of these two age categories seems to have a seasonal pattern [20]. This is likely due to seasonal flu, which affects the younger age categories less. In fact, the "0 to 44 years" and "45 to 64 years' ' categories exhibit a linear trend with respect to time. The "0 to 44 years' ' age category is excluded because this age category has exceptionally low weekly death counts, often resulting in the majority of the data being zeros in less populated provinces. This would not provide a clear trend for weekly death count as the majority of the values cluster around zero. In the case of the territories and Prince Edward Island, the low population value still results in a large amount of zeros in the data even in the "45 to 64 years" categories. There is also no data on Yukon after 2015. This again makes it hard to analyze the actual trend of normalized weekly death count value and therefore the territories and Prince Edward Island are excluded. Quebec is removed after examining Figure 4 as it suggests that it does not share a common trend with other Canadian provinces, which violates the common trend assumption.

Variable selection:

The difference in difference model attempts to estimate the average treatment effect of COVID-19 on provincial weekly death count by calculating the difference in weekly death count trend between the provinces within and outside of the Atlantic Bubble. In order to account for time invariant unobserved provincial characteristics, provincial fixed effect will be included in the model. Since the dataset takes the form of time series data, time fixed effects will also be included to control for unobserved time shocks. Lastly, a estimate for the average treatment effect will be included in conjunction with the product of two dummy variables. The dummy variables are used to indicate whether the province is in the treatment group, and whether if the specific week is after treatment. The product of the dummies indicates whether the province have received treatment. The difference in difference model that is constructed from the above variables takes the form of:

$$Y_{it} = \alpha + \beta_i + \gamma_t + \sigma * (treat_i * post_t) + \epsilon_{i,t}$$

This model aims to explain the percentage deviation from the mean for the weekly death count value given a specific province and time. It takes in the variable β_i which is the province fixed effect with i representing the province, as well as the time fixed effect with t representing the date of the last day of the week. The estimate σ represents the average treatment effect of COVID-19 on percentage change in the mean of weekly death count. Both $treat_i$ and $post_t$ are dummy variables that take the value 0 or 1. The variable $treat_i$ describes whether the province is in the treatment group, provinces within the Atlantic Bubble will have the value of the $treat_i$ variable set to 0 and other provinces will have value of 1 for the $treat_i$ variable. The variable $post_t$ describes whether the date is after January 27, 2020. Since that is the date of the first confirmed case in Canada, the dates after January 27, 2020 will have the $post_t$ variable as 1. The products of the two provide information on whether the province has experienced the treatment intervention. Since the weekly death count in each province is not independent and is serially correlated, this difference in difference model would utilize a robust standard error. This will be done through clustered standard error. In this model, residuals will be clustered by provinces, which will allow the model to account for the serial correlation of residuals in each province [27].

Results

Table 2: Difference in Difference with Cluster Standard Error

Variable	Value
Average Treatment Effect	0.08758
Standard Error	0.028
N	4592

Multiple R-squared(proj model): 0.004693 Adjusted R-squared: -0.1395
F-statistic(proj model): 10 on 1 and 7 DF, p-value: 0.01586

Table 2 displays the coefficient for the average treatment effect of COVID-19 in Canada. The coefficient of the average treatment effect is 0.08785 which can be interpreted as an 8.785 percentage increase from the mean of the weekly death count. The cluster standard error has a value of 0.032 with a t-value of 2.746. The p-value is smaller than 0.05 and the average treatment effect is statistically significant under 95% confidence. A positive value of the estimate suggests that provinces that are affected or "treated" with COVID-19 sees an increase in weekly death count.

The next step will be to perform an analysis to determine whether this result is appropriate to draw causal inference from the result of this model. According to Cunningham (2020), there are four major threats to validity in the difference in difference model [25]. The first is non-parallel trends. The average treatment effect is calculated based on the difference between the differences between the weekly death count trends. In other words, the difference between trends must remain the same without intervention for the calculation of the average treatment effect to be valid. This assumption has been evaluated by using logical reasoning as well visuzaltion shown in Figure 4. The assumption is justified by the similarity of living standards as well as regulatory standards within Canada, which should result in the similarity in mortality rate within Canada. Hence, the similarity in the weekly death count trends. After examining the data visualized in Figure 4, it is confirmed that all provinces in Canada (excluding Prince Edward Island due to lack of data) with the exception of Quebec have very similar trends in weekly death count.

The second threat to validity is compositional differences, which is a concern that arises when working with repeated cross sections. A common concern is that the treatment effect actually originates from the aging sample. This is not totally the case within this dataset as age is separated into four categories, however the demography within each age category might differ over time. As seen from Figure 1, older populations are the drivers of weekly death count in Canada. If the mean age within an age group is changing, that might contribute to part of the death count trend. According to the CIA World Factbook 2010 [37], the distribution of population between the age of 25 and 65 is approixmate uniformly distributed. This means that the age category used in this analysis, which is the 45 to 64 years old category, is approximately uniformly distributed throughout the duration of our analysis. Since the distribution of population within that age category is unchanged, compositional differences should only have a very minor effect on the weekly death count trend.

The third threat is long term effect vs reliability. There is a trade off when it comes to the length of analysis. If the time period evaluated is shorter, the parallel trend assumption is more likely to be satisfied as there are less opportunities for other factors to disrupt the trend. The period used to evaluate the parallel trend assumption spans for 10 years; bias should not be a concern when it comes to selection of the time period for pre-treatment. However, there is only approximately 1 year of span in the post treatment period. This is due to the fact that vaccination efforts began approximately one year after the first confirmed case in Canada. It would not be reasonable to extend the period past January 2021 without accounting for the effect of vaccination. The analysis on long term effects is limited in this report.

The last threat is functional form dependence. The difference in functional form might account for some aspect of the result. The model used in this report assumed that the function forms are linear. From Figure 1, it is shown that the weekly death count trend for the age category selected appears to be linear.

All analysis in this report was conducted in `R version 4.1.1`. Figures were produced using the package ggplot2 (Wickham, 2009). The difference in difference model is produced by the lfe package to derive the average treatment effect using a linear model with multiple group fixed effects and time fixed effects in this section [35][38].

Conclusions

The goal of this report is to determine whether COVID-19 has an impact on the trend of weekly death counts within Canada. This is achieved by using a difference in difference model. Since part of the country of Canada, namely the provinces within the Atlantic Bubble, have a prolonged period of 0 cases, the similarities and the lack of influence from COVID-19 make the Eastern Provinces ideal candidates to act as the control group for the difference in difference approach. The model used to analyze the difference between the trend of Eastern Provinces and non-Eastern Provinces is the following:

$$Y_{it} = \alpha + \beta_i + \gamma_t + \sigma * (treat_i * post_t) + \epsilon_{i,t}$$

Where the response variable is the percentage change from the mean weekly death rate with subscript i representing the provinces and t representing the time. α is the fixed baseline intercept, B is the province fixed effect, γ_t is the time effect. γ represents the average treatment effect. the varaible $treat_i$ is a dummy variable that indicates whether the province is in the treatment group, with Eastern Provinces being assigned 1 and non-Eastern Provinces assigned 0. The $post_t$ variable describes whether if COVID-19 has been introduced in Canada, date is selected to be January 27, 2020 since that is the first official confirmed case of COVID-19 in Canada. The product of the $treat_i$ and $post_t$ variable describes whether the province has experienced the treatment intervention.

It is hypothesized that COVID-19 will result in a positive increase in the weekly death count trend as it is a more deadly variant of the coronavirus [23]. This is later confirmed by Table 2, which shows the result of the difference in difference model and displays the positive coefficient of the average treatment effect which is statistically significant under 95% confidence.

The goal of this report and the choice of model is to form causal inference on the effect of COVID-19 on weekly death count trends. Although this model might have avoided the four threats to validity of the difference in difference model, causal inference still cannot be formed because of confounders that are not accounted for in the model. Causal inference can only be formed once those confounders are accounted for.

The goal of this report and the choice of model is to form causal inference on the effect of COVID-19 on weekly death count trends. Although this model has avoided the four threats to validity of the difference in difference model, there are a set of major possible confounders that this report has not addressed. The possible confounders are the change in the trends of other causes of death. If the increase in death is a result of another cause, such as the seasonal flu, then the increase in death count cannot be attributed to the introduction of COVID-19 virus. This concern can be addressed using the "Leading causes of death, total population, by age group1" dataset from StatCan [33]. The data is visualized below:

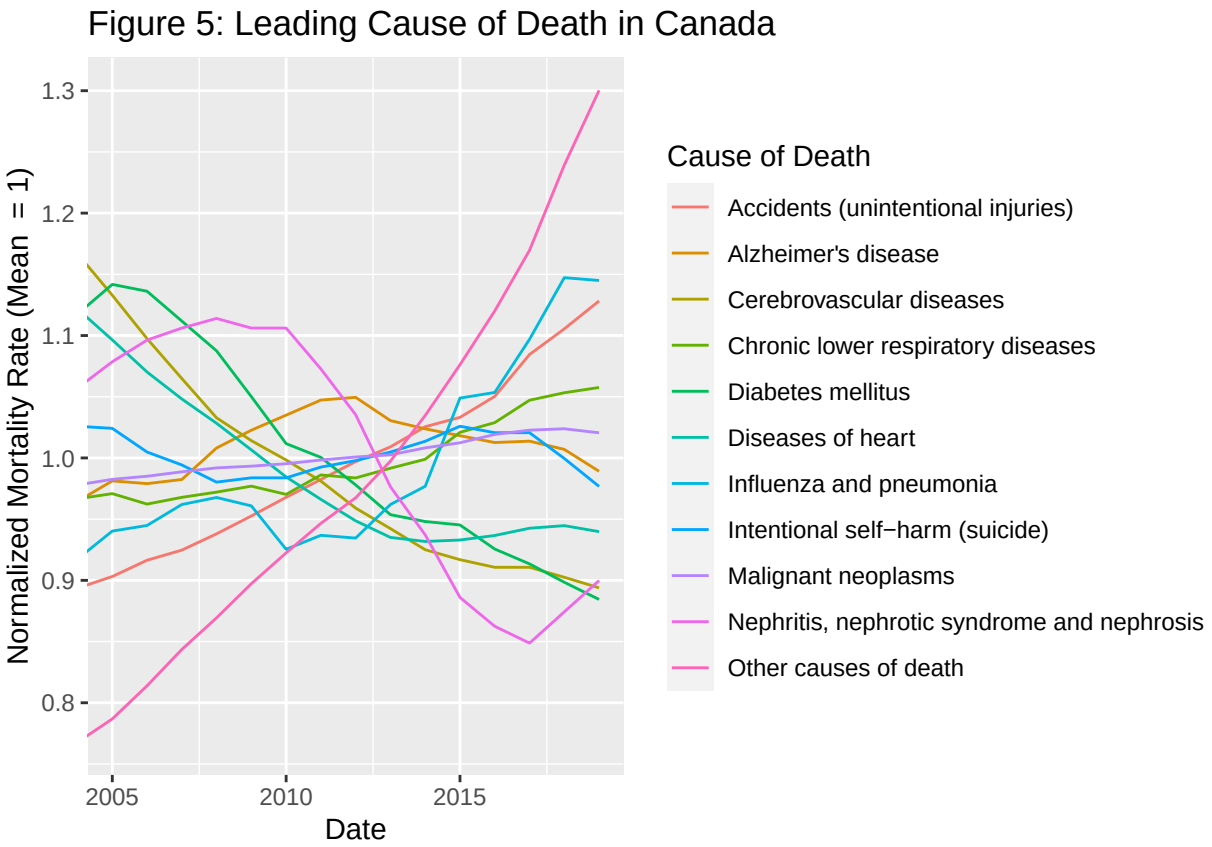


Figure 5 depicts the normalized mortality rate of the top 10 leading causes of death in Canada. Figure 5 also addresses one of the suspected confounders, which is the seasonal flu. According to Figure 1, the weekly death count seems to have a seasonal pattern even before the introduction of the COVID-19 virus, which is attributed to the seasonal flu [20]. However, it is shown in the figure above that death caused by influenza has a steady trend. According to this figure, the trends of leading cause of death seems to have a steady trend which would be captured in the first degree of difference calculation in the Difference in Difference 12 model. There is significant evidence towards a causal relationship between COVID-19 and the increase in death count across Canada.

Weaknesses and Next Steps

One confounder is the difference in lockdown policies [24]. Since lockdown measurements are determined by the provincial governments, it is highly likely that it differs across provinces. Some variables that describe the strictness of lockdown policies should be included in the model to account for the difference of lockdown policies across provinces.

The dataset used in this report is a provisional dataset; there is no finalized dataset that is nearly as updated. Since the dataset is provisional, it might not reflect all the deaths that occurred during a specific time period or might even be updated in the future. There is also the issue with lack of information in specific provinces due to their low population. With a low population the weekly death count for individuals between the age of 45 and 64 are often 0, resulting in sparse data which make analysis of trends difficult or might even skew the result. In this report, those provinces are removed but might cause the results to be biased. There also exists the question of when COVID-19 was actually introduced in Canada. The seriousness of COVID-19 was significantly underestimated at the beginning of its outbreak [22]. This results in a lack of implementation of the testing procedures for COVID-19. The first confirmed case of COVID-19 in Canada was January 27, 2020, but that does not mean that is the first case of COVID-19 in Canada. According to Figure 4, there

is an unexplained increase in death rate before January 2020, further research is needed to determine the exact date. One last issue with this dataset is the weekly death count is not shown as a proportion to a fixed number of population, meaning that this dataset describes the actual number of weekly deaths but not a ratio. In this report, the weekly death count value is normalized by the provincial mean to account for the difference in provincial population. However this would fail to account for population growth. The next step would be to acquire population data and acquire a ratio accordingly instead of normalizing by the historic mean.

Bibliography

1. Statistics Canada. (2022). Provisional weekly death counts, by age group and sex. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310076801

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. Loomba, S., de Figueiredo, A., Piatek, S.J. et al. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 5, 337–348 (2021). https://doi.org/10.1038/s41562-021-01056-1

5. Greene CM, Murphy G. Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation. J Exp Psychol Appl. 2021 Jun 10. doi: 10.1037/xap0000371. Epub ahead of print. PMID: 34110860.

6. McKinsey & Company. (2021, December 2). The coronavirus effect on global economic sentiment. McKinsey & Company. Retrieved December 13, 2021, from https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-coronavirus-effect-on-global-economic-sentiment.

7. Perera, M. (n.d.). Gen Z: Government ethics during this time of pandemic. Markkula Center for Applied Ethics. Retrieved December 13, 2021, from https://www.scu.edu/ethics-spotlight/covid-19-ethics-health-and-moving-forward/top-issues-related-to-government-ethics-during-this-time-of-pandemic/.

8. Shearer, E., & Gottfried, J. (2020, August 27). News use across social media platforms 2017. Pew Research Center's Journalism Project. Retrieved December 13, 2021, from https://www.pewresearch.org/journalism/2017/09/07/news-use-across-social-media-platforms-2017/.

9. The Regulatory Review. (2021, August 23). Responding to deepfakes and disinformation. The Regulatory Review. Retrieved December 13, 2021, from https://www.theregreview.org/2021/08/14/saturday-seminar-responding-deepfakes-disinformation/.

10. Thomson Reuters. (2020, October 6). Facebook, twitter take action over Trump's misleading COVID-19 posts. Reuters. Retrieved December 14, 2021, from https://www.reuters.com/article/us-twitter-trump-idUSKBN26R2Z3.

11. Thomson Reuters. (2020, December 16). Fact check: Covid-19 is not a seasonal flu. Reuters. Retrieved December 14, 2021, from https://www.reuters.com/article/uk-factcheck-seasonal-flu-idUSKBN28Q1LC.

12. Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021, February 5). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nature News. Retrieved December 14, 2021, from https://www.nature.com/articles/s41562-021-01056-1.

13. Young, L. (2021, September 29). Unvaccinated 60 times more likely to end up in ICU with COVID-19, Ontario Data shows. Global News. Retrieved December 20, 2021, from https://globalnews.ca/news/8230051/covid-vaccine-hospitalization-risk-ontario/

14. Global News. (2021, July 17). Here's what provinces are planning for covid-19 reopening across Canada - national. Global News. Retrieved December 20, 2021, from https://globalnews.ca/news/8036166/covid-reopening-plans-provinces-july-2021/

15. Government of Canada. (2011, October 20). Government of Canada. Canada.ca. Retrieved December 20, 2021, from https://www.canada.ca/en/health-canada/services/health-care-system/canada-health-care-system-medicare/canada-health-act-frequently-asked-questions.html#a3

16. Life satisfaction. The Conference Board of Canada. (n.d.). Retrieved December 20, 2021, from https://www.conferenceboard.ca/hcp/provincial/society/life-satisfaction.aspx

17. Government of Canada. (2021, February 16). Government of Canada. Canada.ca. Retrieved December 20, 2021, from https://www.canada.ca/en/employment-social-development/programs/seniors-action-report.html

18. MacDonald, M. (2021, May 1). The covid-zero approach: Why Atlantic Canada excels at slowing the spread of covid-19. Atlantic. Retrieved December 20, 2021, from https://atlantic.ctvnews.ca/the-covid-zero-approach-why-atlantic-canada-excels-at-slowing-the-spread-of-covid-19-1.5410217

19. Global News. (2020, March 7). Coronavirus: Here's a timeline of covid-19 cases in Canada - national. Global News. Retrieved December 20, 2021, from https://globalnews.ca/news/6627505/coronavirus-covid-canada-timeline/

20. Government of Canada. (2021, March 25). Government of Canada. Canada.ca. Retrieved December 20, 2021, from https://www.canada.ca/en/public-health/services/publications/diseases-conditions/fluwatch/2019-2020/annual-report.html

21. Global News. (2020, March 7). Coronavirus: Here's a timeline of covid-19 cases in Canada - national. Global News. Retrieved December 20, 2021, from https://globalnews.ca/news/6627505/coronavirus-covid-canada-timeline/

22. Sparrow, A. (2021, March 18). The Chinese government's cover-up killed Health Care Workers Worldwide. Foreign Policy. Retrieved December 20, 2021, from https://foreignpolicy.com/2021/03/18/china-covid-19-killed-health-care-workers-worldwide/

23. Coronavirus: the science explained - UKRI. (n.d.). What is coronavirus? the different types of coronaviruses. UKRI. Retrieved December 20, 2021, from https://coronavirusexplained.ukri.org/en/article/cad0003/

24. Paisley Sim: Covid-19 policy stringency across provinces. Max Bell School of Public Policy. (2021, May 21). Retrieved December 20, 2021, from https://www.mcgill.ca/maxbellschool/article/articles-max-policy/covid-19-policy-stringency-across-provinces

25. Cunningham, S. (n.d.). Causal inference: The mixtape. in. Retrieved December 20, 2021, from https://mixtape.scunning.com/difference-in-differences.html

26. Government of Canada. (2021, September 1). Where our legal system comes from. About Canada's System of Justice. Retrieved December 20, 2021, from https://www.justice.gc.ca/eng/csj-sjc/just/03.html

27. Kuriwaki, S. (2021, December 19). Difference-in-differences estimation in R (2 of 2). Vimeo. Retrieved December 20, 2021, from https://vimeo.com/409267190

28. Government of Canada, S. C. (2021, June 15). Canadian Vital Death Statistics Database (CVSD) linked to Discharge Abstract Database (DAD) and National Ambulatory Care Reporting System (NACRS). Government of Canada, Statistics Canada. Retrieved December 20, 2021, from https://www.statcan.gc.ca/en/microdata/data-centres/data/cvsd-nacrs

29. Katella, K. (2021, December 17). Comparing the COVID-19 vaccines: How are they different? Yale Medicine. Retrieved December 20, 2021, from https://www.yalemedicine.org/news/covid-19-vaccine-comparison

30. Aiello, R., & Forani, J. (2020, December 14). 'V-day': First covid-19 vaccines administered in Canada. Coronavirus. Retrieved December 20, 2021, from https://www.ctvnews.ca/health/coronavirus/v-day-first-covid-19-vaccines-administered-in-canada-1.5230184

31. Communications Nova Scotia. (2021, December 17). Travel. Coronavirus (COVID-19). Retrieved December 20, 2021, from https://novascotia.ca/coronavirus/travel/

32. Jacobs, E. (2020, December 12). Canadians in 'atlantic bubble' take drastic measures to keep infections low. NPR. Retrieved December 20, 2021, from https://www.npr.org/2020/12/12/945896654/canadians-in-atlantic-bubble-take-drastic-measures-to-keep-infections-low

33. Statistics Canada. (2022). Leading causes of death, total population, by age group1. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310039401

34. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

35. Gaure S (2013). "lfe: Linear group fixed effects." The R Journal, 5(2), 104-117. User documentation of the 'lfe' package, https://journal.r-project.org/archive/2013/RJ-2013-031/RJ-2013-031.pdf.

36. Warin T, & Romain L. D. (2019). "statcanR: Client for Statistics Canada's Open Economic Data" doi: 10.6084/m9.figshare.10544735.v1

37. The World Factbook 2010. Google Books. Retrieved December 20, 2021, from https://books.google.ca/books?id=5lc9jxSKfuEC&q=world%2Bfactbook&pg=PP2&redir_esc=y#v=snippet&q=world%20factbook&f=false

38. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

39. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

Appendix

A1: Ethics Statement

a concern when data is required. The dataset used in this report is acquired from Statistics Canada. It is an open dataset published by a government agency. In research, reproducibility is key, and without reproducibility, the result of the research cannot be verified. In this report, the method of data acquisition, cleaning, as well as method is clearly explained with the aim to aid reproducibility. Another significant issue that is ongoing in the research field is p-hacking. All the information utilized to justify decisions are properly cited. The use of existing sources also aims to provide quantitative information instead of qualitative information, this would not leaves the researcher room to make discretionary decisions. The method used to determine the treatment intervention date is an example. The treatment intervention date is the date of the first confirmed COVID-19 case in Canada, and the information is acquired from reputable news publishers.