Investigation of the relationship between vaccines and fatal COVID cases

Assignment 2

Sagnik Adusumilli - 1003157567

Introduction

The COVID-19 pandemic has been around for almost over a year now, and over 3,500 people have died in Toronto so far[1]. Trusted sources like the Center for Disease Control (CDC) and World Health Organization (WHO)[2][3] all claim that getting vaccinated is an effective measure for protecting oneself and ending the pandemic. However there are many there are still some instances of vaccine hesitancy that remain[4]. WHO defines vaccines hesitancy as: "delay in acceptance or refusal of vaccines despite the availability of vaccines"[5].

In this paper, using regression analysis, I will investigate if the vaccine was effective in decreasing the number of fatalities caused by COVID and also examine how it might effect the number of fatal COVID cases for different genders and different age groups. I think the following results might be observed:

- Given the wide support of vaccines from trusted sources, we might observe that the odds of fatal cases decrease with time after the introduction of vaccines.
- There might be more fatalities observed in the older age bands since older people tend to have a weaker immunity[15]. Perhaps the difference in number of fatalities in each age group might decrease after vaccinations have been introduced.
- I think that the vaccine would have similar levels of effectiveness on each gender. I did not come across any article that claimed the contrary.

Data

Data Collection Process

The dataset can be found on the Toronto Open data portal[6]. The Data was extracted from the provincial Case & Contact Management System (CCM)[7]. These are the cases that were reported to Toronto Public Health[8]. This dataset is refreshed every week and new additional cases are reported. This ensures that the analysis is not done on outdated data. However, this also means that certain fields could be added or removed over time. This makes the the data source a little unstable. However the attributes that are being investigated (Gender, age and time) will most likely be constant. Furthermore, this also implies that that the findings of this report may not be accurate as more data is entered. However, the results can be updated accordingly by following the same methodology.

Data Summary

The dataset records over 180,000 cases that were reported to Toronto Public health since January 2020. Each row in the table is a case that that was reported and there are 18 columns that record different information pertaining to the case. For our investigative purposes, we will only consider the following attributes.

- Reported Date: The date on which the case was reported to Toronto Public Health.
- Outcome: This is a categorical variable that indicates the outcome of the case.
  - FATAL: Cases with a fatal outcome
  - RESOLVED: Non-fatal cases where either the person recovered or reported date is more than 14 days from symptom onset and the person is not currently hospitalized
  - ACTIVE: All cases that are not in the above two categories
- Age Group: Age at the time of the illness. This is also a categorical variable that divides the age into the following bands by years: ≤ 19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+, unknown (This is reported as blank)
- Client Gender: The gender reported by the person themselves

Data Cleaning

- First all rows with missing values were removed. After removing the missing values, the dataset contained over 175,000 rows. This was a large sample size hence there was no need to estimate the missing values.
- Then the variables were renamed for convenience and simplicity
  - 'Reported' Date was renamed to be 'Date'
  - 'Client Gender' was renamed to be 'Gender'
- For the purpose of this study, the only outcome of interest is to know if a COVID case is fatal or not. Hence all non-fatal outcomes were recorded to be 0 and all fatal outcomes were coded as 1. This transformed the variable to a binary variable that will be used as the response variable for the regression model. The variable was renamed to 'Fatal' to denote its new meaning.
- Initially the Gender variable contained several genders, however 'PREFER NOT TO SAY', 'NON LISTED, 'PLEASE SPECIFY', 'UNKNOWN' means the same thing semantically. Hence these values were all grouped under 'UNKNOWN'.
- Created a new categorical variable named 'Period' to indicate if a case occurred after vaccinations were made available to the public or before. This was done by comparing each reported date of a case to '2020-12-14'[10], which is the date that vaccinations were introduced.
- Lastly, a new variable was created to represent the progression of time. The 'Date' variable contains individual dates which does not directly indicate how much time has passed and also cannot be used as input directly for regression analysis. Hence a new numerical variable 'Day' was created. For each row (This represents a single COVID case) the the 'Day' value is the number of days that have passed since the first COVID case was recorded by Toronto Public Health. This first day is set to 0.

Numerical Summaries

Table 1: Demographic breakdown by Gender

Reported Gender	Count
MAN/BOY	63
NON-BINARY	25
OTHER	22
TRANS MAN	3
TRANS WOMAN	3
TRANSGENDER	21
UNKNOWN	175034
WOMAN/GIRL	64

From this table we can see that for most of the cases we don't know the gender of the person. Therefore gender may not be a significant variable to analyze.

Table 2: Demographic breakdown by Age Group

Age Group	Count
19 and younger	26213
20 to 29 Years	35963
30 to 39 Years	30502
40 to 49 Years	25217
50 to 59 Years	24823
60 to 69 Years	15682
70 to 79 Years	7777
80 to 89 Years	5869
90 and older	3189

Looking at this table, it seems that most of the COVID cases reported are from people who fall in the age band 20 to 39 years. This maybe because people of this age band mostly work full time and therefore may have more difficulty in avoiding social contact.

For the following numerical summaries it should be noted that we are only considering COVID case reports with no missing values. The first COVID case was recorded on 2020-01-23 and the last recorded date of a COVID case was 2021-10-19 which means 635 days have passed since the first COVID case was recorded. During that time, the minimum number of cases recorded on a day is 1 which were recorded on the following days 2020-02-04, 2020-02-21, 2020-02-25, 2020-02-26, 2020-02-27, 2020-02-28, 2020-02-29, 2020-03-02, 2020-03-03

The most number of cases recorded in a day is 1622 which occurred on 2021-01-07. The total average number of cases in Toronto for the dataset is 289.1666667 per day

Data plots

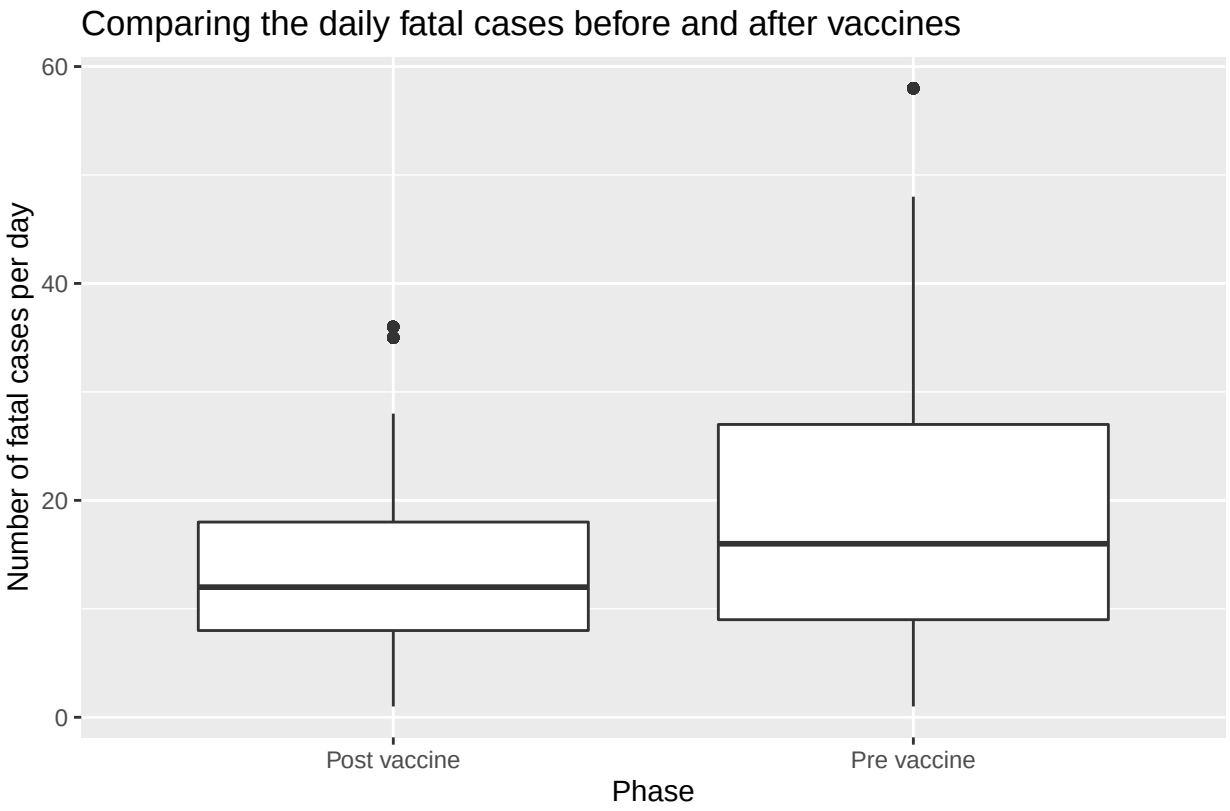

Figure 1

The box-plot labeled 'Post vaccine' represents the fatal cases after vaccination was made available to Ontarians. The other box-plot represents the number of fatal cases before vaccinations. If we compare the IQR range and the max values of each box-plot, we can see that the 'Post vaccine' box-plot has smaller values. This indicates that there were fewer number of daily fatal cases on average after vaccinations were made accessible and furthermore the smaller size also suggests that there was less variability in the number of cases as well.

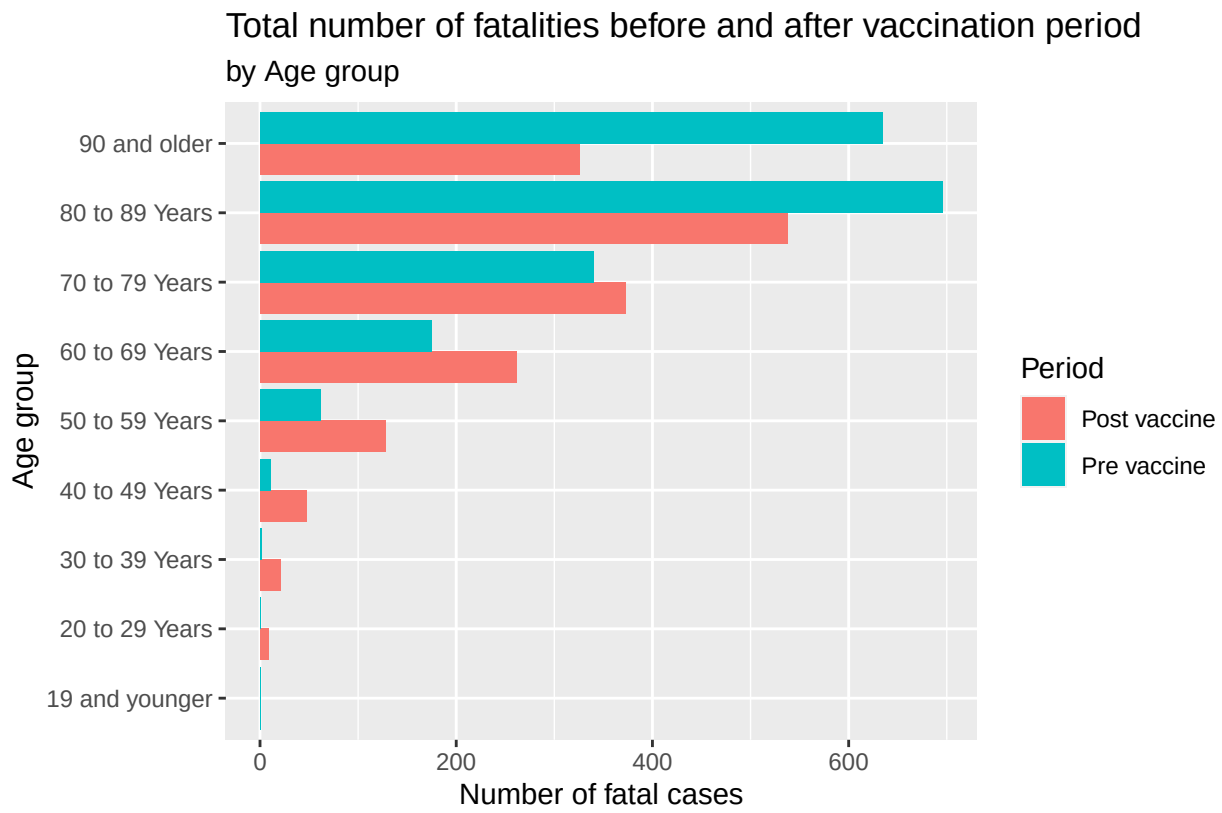

Figure 2

Each bar in the bar graph represents the total number of fatal cases per age group. The red bars are fatal cases that happened before the vaccine was introduced. The teal bars represent the number of fatal cases that happened after the vaccines was introduced.

Going by bar lengths for each age group, it would seem that the number of fatalities increase for older age groups. We can also see that the number of fatal cases for people who are 80 and older were much higher before the vaccine was introduced. However, it seems that for all other age groups, more fatalities occurred after the vaccine was introduced.

This does seem to contradict the conclusion drawn from the previous graph. However, there can be several factors that could explain this. Firstly we must take into account that the vaccines were not made available to all the age groups at once. Older people and other people classified as “at-risk” were given access to the vaccines first [11]. Even after the vaccinations were introduce, there were several new variations that had higher infection rates [12]. These factors could explain why fatal cases recorded for the younger groups may have been higher in this period. It should also be noted that the number of days before vaccine is less than the number of days after vaccine. So the latter period had more time to accumulate fatal cases.

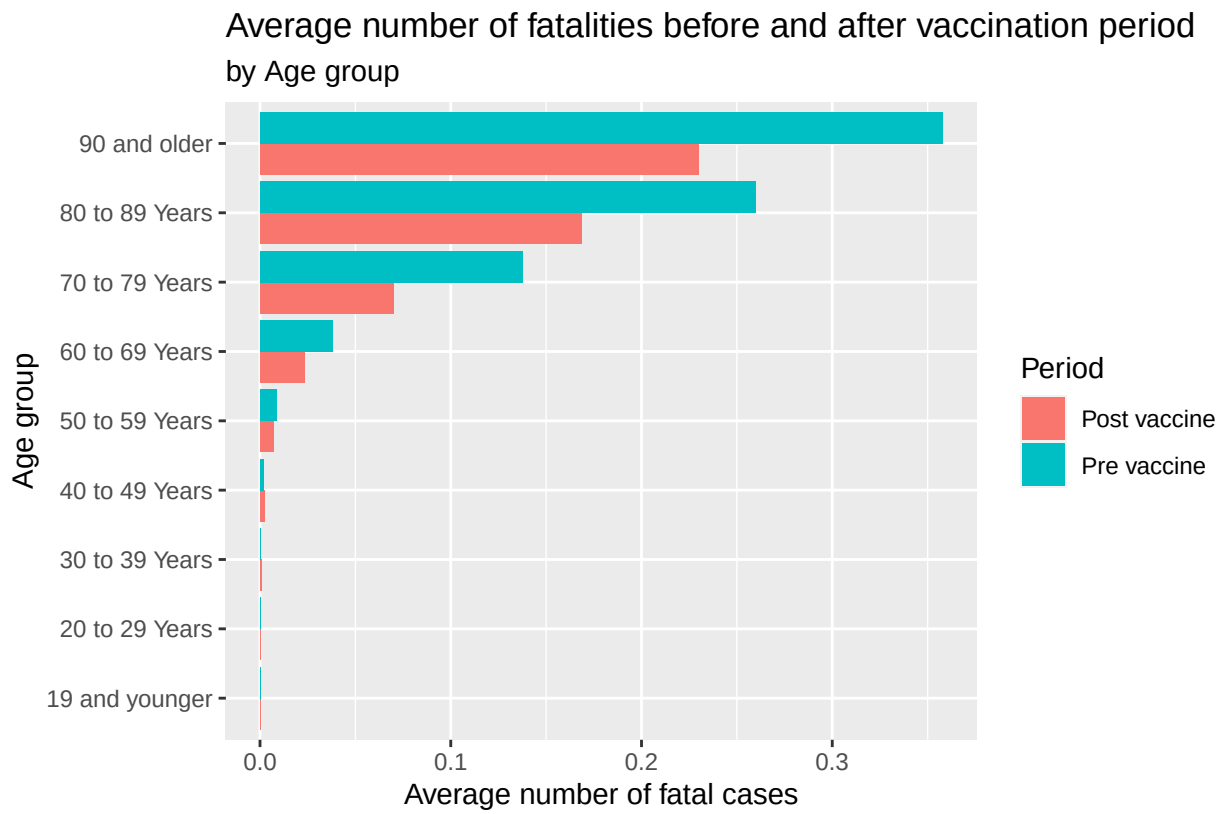

Figure 3

This graph compares the average number of fatal cases before and after vaccination was introduced for each age group. We can see that average number of fatal cases decreased for every age group above 50. The other age groups had a very small average from the start so the vaccination did not effect those ratios. We could also conclude that the vaccination did not increase the likelihood of fatal cases, as the previous graph would imply for some age groups.

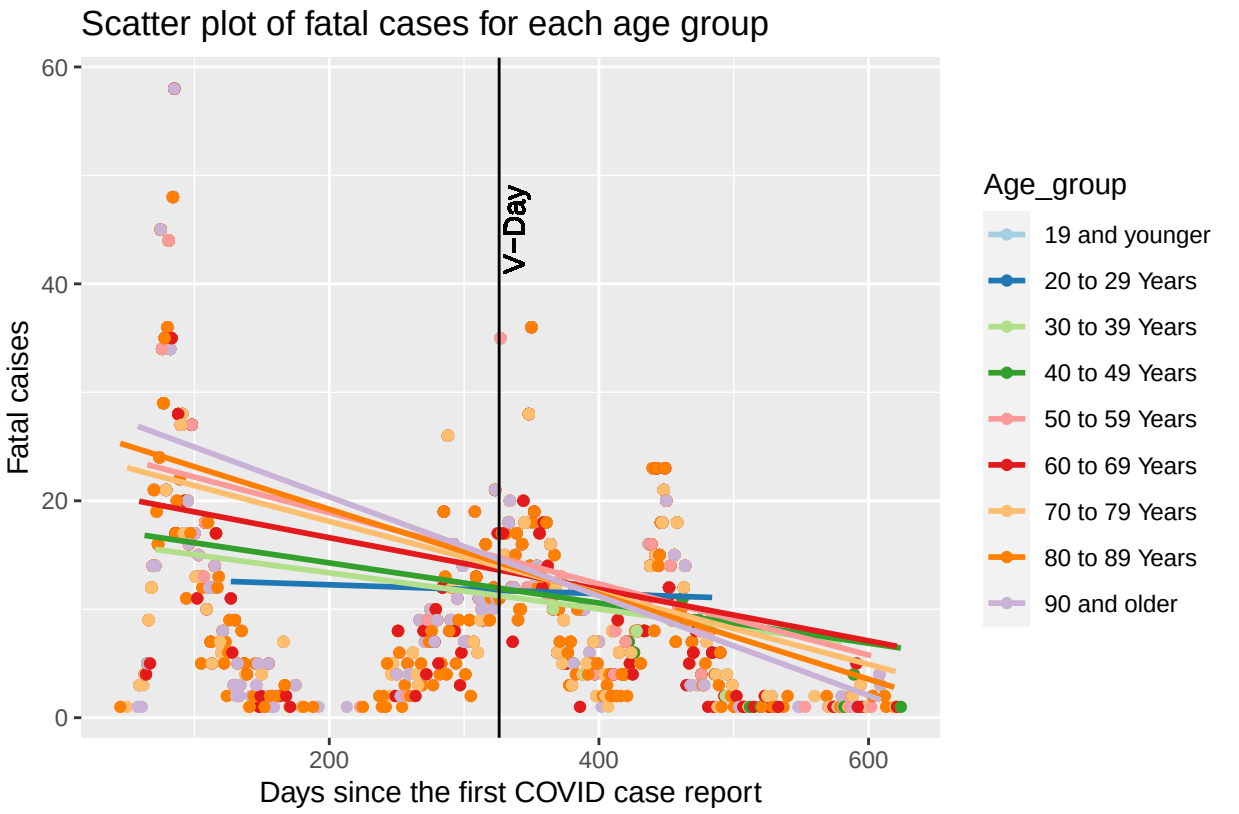

Figure 4

This graph adds time as a factor to show how the number of fatal cases changed for each age group over time. Regression lines were fitted to show the trend for each age group. "V-day" marks the day vaccination was made accessible to the public. From this we can see that the Vaccines were introduced during the 2nd wave of COVID, so the number of the fatalities would still be expected to rise. It should also be noted that the amplitude of the subsequent waves after "V-day" are significantly lower than the amplitude of the wave before it. Secondly, we can also see that the trend line for all age groups have a decreasing slope with respect to the number of days. This means that the average number of fatal cases decreased for every age group. This supports the hypothesis that the vaccines help decrease the odds of fatalities for all age groups.

Note: I am not doing any graphical summaries on gender because as shown in the numerical summaries section, for most of the cases, the gender was unknown. So I do not think that there will be any useful insights we could gain.

All analysis for this report was programmed using `R version 4.1.1`.

Methods

Before continuing with the investigation, I will explain the model which will be used for the analysis in this section. The model that will be used is a binary classifier called logistic regression. It works as follows. Suppose we want to predict binary outcome variable Y based on the predictors $X_1, X_2, \ldots, X_k$. Here we assume that Y follows a Bernoulli distribution where $P(Y)$ is the probability of outcome Y occurring. Then Logistic regression allows us to express Y as a function of X_1 and X_2 as the following formula:

$$log\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

The term $log\left(\frac{P(Y)}{1-P(Y)}\right)$ is known as the log-odds $P(Y)$ is the probability of outcome Y occurring. β_0 represents the default value of the log-odds. The coefficient B_i for $i \in \{1, 2, \ldots k\}$represent the change in log odds for one unit change in X_i assuming that all other predictor values being constant. ϵ is a random variable that has a standard normal distribution. This term represents the error term in our prediction.

$$log\left(\frac{P(\hat{Y})}{1-P(\hat{Y})}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$$

To get a prediction from this model we can do the following process: Let $\eta = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$. Now we have that $log\left(\frac{P(\hat{Y})}{1-P(\hat{Y})}\right) = \eta$. Therefore we get that $P(\hat{Y}) = \frac{e^\eta}{1+e^\eta}$. we now can get an estimated probability of outcome Y from predictors. Now we can set a threshold γ where if $P(\hat{Y}) \geq \gamma$ we predict 1 and 0 otherwise. Usually the γ is set to 0.5 which means 50% probability

For each estimated coefficient we also compute the P-value. This P-value of coefficient β_i is the probability that we got an estimate $\hat{\beta}_i$ under the assumption that the true parameter value β_i is 0. If the probability of getting the value $\hat{\beta}_i$ under this assumption is less or equal to significance level α (usually 0.05 in most cases), we reject the assumption (This is the null hypothesis) and conclude that $\beta_i \neq 0$. We can also use the P-value to compute the confidence interval (CI) of each estimate. The 95% CI [a, b] implies that if we were to run the gradient descent algorithm several times, then 95% of the time the estimate $\hat{\beta}_i$ will be in between a and b.

Now, I will explain how I will use this model in my investigation. First I will construct a logistic model using the full dataset and check the significance level for each predictor. If the P-values are low, then it would imply that these predictors are statistically significant for predicting the Fatality of a COVID case. I will disregard predictors with high P-values. My hypothesis however, is that "Age_group" and "Day" are significant predictors since the graphs showed that these variables have relationship with the fatality of the cases.

Now to study the effect of the vaccines on the significant variables, I will construct two models on different datasets. The first model will be constructed with the dataset containing only COVID cases that happened before vaccination was made available to the public. The second model will be constructed with the dataset containing on the cases that happened after that day. Then I will compare the coefficients and the P-value for each predictor in the two models. A big difference in the coefficients of a predictor or a large change in P-value would imply that the introduction of vaccine has effected the relationship of the variables and the outcome, which is the Response variable 'FATAL' in our case.

Results

For the initial model, the following variables was used:

- FATAL: this is the response variable. It is a binary categorical variable where the value 1 denotes that a case was fatal and 0 denotes a non-fatal case
- Age_Group: This is a categorical predictor variable where each age band is a category
- Gender: This is a categorical predictor variable where each gender is a category
- Day: This a numerical predictor variable where the day value is the number of days that passed since the first COVID case was recorded by Toronto Public Health

Table 3: full model output

	Coefficient estimate	P-value
(Intercept)	-18.6006151	0.9083415
Day	-0.0027158	0.0000000
as.factor(Gender)NON-BINARY	1.9486919	0.9951968
as.factor(Gender)OTHER	-1.3745299	0.9964411
as.factor(Gender)TRANS MAN	2.5391181	0.9974327
as.factor(Gender)TRANS WOMAN	0.5523743	0.9994481
as.factor(Gender)TRANSGENDER	-0.5536723	0.9986264
as.factor(Gender)UNKNOWN	9.4716180	0.9532493
as.factor(Gender)WOMAN/GIRL	0.6572461	0.9976845
as.factor(Age_group)20 to 29 Years	1.9282240	0.0660006
as.factor(Age_group)30 to 39 Years	2.9269086	0.0041680
as.factor(Age_group)40 to 49 Years	4.0205956	0.0000671
as.factor(Age_group)50 to 59 Years	5.2061533	0.0000002
as.factor(Age_group)60 to 69 Years	6.5093776	0.0000000
as.factor(Age_group)70 to 79 Years	7.7424646	0.0000000
as.factor(Age_group)80 to 89 Years	8.5863900	0.0000000
as.factor(Age_group)90 and older	8.9863275	0.0000000

For regression, all variables must be converted to numerical values. Hence there were additional predictors to represent categories in categorical variables. For example suppose an input has the gender "TRANS MAN". In that case, all other gender related variables will be set to 0 and the variable "TRANS MAN" will be set to 1 to indicate the gender. However, it should be noted that there is one variable missing for each category. For example, for gender the variable to represent "MAN/BOY" is missing. This is because the model treats this as the reference category. This means that if all gender related variables is set to 0, the default assumption of the gender is "MAN/BOY". I will explain how to interpret the coefficient estimates in later models. This model is only used to filter the variables based on P-values. We see that for all gender related variables, they have very high P-values. This means that under the null hypothesis (the real coefficient value is 0), there is a high probability that we can derive this estimate. This implies that these factors are not useful in predicting the fatality of a case and hence should be left out of the model. In other words, it would seem that all genders are similar odds of getting a fatal COVID case. This was an expected result since Table 1 shows that the most of the cases had unknown gender.

After removing gender from the list of predictors, I constructed two models on different datasets to examine how the relationships changed between the variables after vaccination was introduced. The first dataset contained COVID cases after vaccination was introduced to the public and the second dataset contained COVID cases after this date

Table 4: Model ouput on dataset before vaccinations

	Coefficient estimate	P-value
(Intercept)	-7.6354444	0.0000000
Day	-0.0040544	0.0000000
as.factor(Age_group)20 to 29 Years	-0.6025285	0.6700175
as.factor(Age_group)30 to 39 Years	0.2648188	0.8288355
as.factor(Age_group)40 to 49 Years	2.0706479	0.0474702
as.factor(Age_group)50 to 59 Years	3.7587677	0.0001931
as.factor(Age_group)60 to 69 Years	5.2496963	0.0000002
as.factor(Age_group)70 to 79 Years	6.5998446	0.0000000
as.factor(Age_group)80 to 89 Years	7.2821774	0.0000000
as.factor(Age_group)90 and older	7.7199595	0.0000000

This is the model output for model constructed on the dataset containing cases before vaccination were introduced. Even though some age predictors have high P-value there other age predictors that have low P-value and therefore the predictor Age_group is significant as a whole. The P-value of Day variable is also very low hence it is also a statistically significant predictor. Note that none of the P-values are actually 0, but the `kable()` function in R reports very low value as approximately 0.

Now I will discuss the interpretation of the coefficient estimates. First consider the estimate of the Day predictor. This value means that for every one unit increase in day changes the log-odds by -0.0040544. This also means that the odds of getting a fatal covid case for every one unit increase in Day decreases by $(e^{0.0040544}) = 1.0040626$[14]. This is under the assumption that all other predictor values remains constant.

For the categorical variable, notice that the age group "19 years and younger" is not listed. This means that this is the reference category. Every coefficient of the other age category is relative to this category. For example the coefficient 7.7424646 for the "70 to 79 Years" means that the log-odds of a fatal COVID case for this age group is 7.7424646 more than the log-odds of fatal COVID case for "19 and younger age band". This also means that odds of a fatal case for the age group "70 to 79 years" is $(e^{7.7424646}) = 2304.1441846$[14] times the age group "19 years and younger". This is under the assumption that all other predictor value remains constant.

Table 5: Model ouput on dataset after vaccinations

	Coefficient estimate	P-value
(Intercept)	-20.3290076	0.8697944
Day	-0.0005439	0.1851871
as.factor(Age_group)20 to 29 Years	12.5840076	0.9191783
as.factor(Age_group)30 to 39 Years	13.5810618	0.9127991
as.factor(Age_group)40 to 49 Years	14.6006629	0.9062815
as.factor(Age_group)50 to 59 Years	15.6359295	0.8996703
as.factor(Age_group)60 to 69 Years	16.8321065	0.8920402
as.factor(Age_group)70 to 79 Years	17.9718173	0.8847796
as.factor(Age_group)80 to 89 Years	18.9537772	0.8785316
as.factor(Age_group)90 and older	19.3311754	0.8761324

This is the the output of the model constructed on the dataset after vaccinations were made available. The coefficient estimates are much higher for the age group predictors. However, the high P-value indicates that the null hypotheses is most likely true. In other words, the true coefficient values are most likely 0. This means that there is no significant difference between the odds of getting a fatal covid case for each age group.

What is interesting to note is that the P-value for the day variable is also slightly high. This implies that there is less certainty if the number of fatalities decrease as the days pass. If we look at Figure 4, we can see that the number of fatalities since day 500 are very few in number. Hence there is not much room for the number of fatalities to decrease further. In this context the high P-value and low estimate value makes sense.

All analysis for this report was programmed using `R version 4.1.1`. I used the `glm()` function in base `R` to derive the estimates of a frquentist logistic regression in this section.

Conclusions

The aim for this paper was to investigate how effectively vaccines were able to reduce the risk of someone contracting a fatal case of COVID. We examined its effect on demographic groups based on gender and age. We also investigated if the vaccines also caused the fatal cases to do down with time.

Following are the main results from the investigation

- The model constructed on data before vaccines were present showed that the odds of getting a fatal COVID case increased progressively as the age groups got older.
- From figure 3, we saw that the mean number of fatal cases were close to 0 for younger age groups. From this we can infer that these age groups must have low odds of getting a fatal COVID case. We also know the model constructed on data after vaccines were made available suggested that the new odds of getting COVID was similar for each age group. From these two results we could infer that the older age groups would also have low odds of getting a fatal COVID.
- The regression analysis did not show any significant insights between time and the number of fatal COVID cases. However, in figure 4 which shows the number of fatal COVID cases vs days, we saw that for the COVID waves after the vaccines were introduced has a smaller amplitude than the wave before it. This perhaps suggest that the vaccines were able to suppress the number of fatalities that could have occurred.

Overall, The analysis done in this reports supports the hypothesis that vaccines help prevent fatal COVID cases especially among older people.

Weaknesses

- The first major weakness is that we did not have access to vaccination data. In other words we did not know how many of the fatal cases that occurred included fully vaccinated people. If Toronto Public Health also reported if a person was vaccinated, we could have more direct evidence to show a link between vaccines and fatality.
- For most of the fatalities the gender was marked as 'UNKNOWN', hence we could not gain any meaningful insights into the relationship between Gender and fatality
- It should also be noted that a case marked as active could later be fatal if the person dies later. Our analysis did not take this into account and simply marked active cases as non fatal.
- Lastly as was mentioned before, due to the constant updates of the dataset, the results of this analysis may not be as accurate in the future since the data might reveal that the relationships between the variables have changed.

Next Steps

- In the future if we could know what vaccine a patient was given and what variant of COVID infection they had, we could investigate how effective each vaccine is against different COVID strains.
- Due to the constricted time, I did not have time to research the other variables present in the dataset. Perhaps there could a be regression model built to predict the number of cases for each day. This model could use more variables present in the dataset to predict this.
- The regression analysis done on the data was from a frequentist's perspective, which assumes no prior distribution of the outcome variable. Perhaps if a domain expert in the virology could be consulted, they could provide help select a reasonable distribution of the predictor and the outcome to help us construct a bayesian model. This perhaps could give stronger and more accurate results.

Bibliography

1. City of Toronto, "Covid-19: Case counts," City of Toronto, 23-Sep-2021. [Online]. Available: https://www.toronto.ca/home/covid-19/covid-19-pandemic-data/covid-19-weekday-status-of-cases-data/. [Accessed: 26-Oct-2021].

2. "Vaccines for covid-19," Centers for Disease Control and Prevention. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/vaccines/index.html. [Accessed: 26-Oct-2021].

3. "Covid-19 vaccines," World Health Organization. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines. [Accessed: 26-Oct-2021].

4. S. Machingaidze and C. S. Wiysonge, "Understanding covid-19 vaccine hesitancy," Nature News, 16-Jul-2021. [Online]. Available: https://www.nature.com/articles/s41591-021-01459-7. [Accessed: 26-Oct-2021].

5."Vaccine hesitancy: What it means and what we need to know ..." [Online]. Available: https://www.who.int/immunization/research/forums_and_initiatives/1_RButler_VH_Threat_Child_Health_gvirf16.pdf?ua=1. [Accessed: 26-Oct-2021].

6. "Open data dataset," City of Toronto Open Data Portal. [Online]. Available: https://open.toronto.ca/dataset/covid-19-cases-in-toronto/. [Accessed: 26-Oct-2021].

7. Government of Ontario, Ministry of Health and Long-Term Care, "Mers-COV (novel coronavirus) - ministry programs - health care professionals - MOHLTC," Coronavirus - Ministry Programs - Health Care Professionals - MOHLTC. [Online]. Available: https://www.health.gov.on.ca/en/pro/programs/publichealth/coronavirus/. [Accessed: 26-Oct-2021].

8. City of Toronto, "Toronto Public Health," City of Toronto, 21-Feb-2019. [Online]. Available: https://www.toronto.ca/city-government/accountability-operations-customer-service/city-administration/staff-directory-divisions-and-customer-service/toronto-public-health/. [Accessed: 26-Oct-2021].

9. "Opendatatoronto," README. [Online]. Available: https://cran.r-project.org/web/packages/opendatatoronto/readme/README.html. [Accessed: 26-Oct-2021].

10. R. Aiello and J. Forani, "'V-day': First covid-19 vaccines administered in Canada," Coronavirus, 14-Dec-2020. [Online]. Available: https://www.ctvnews.ca/health/coronavirus/v-day-first-covid-19-vaccines-administered-in-canada-1.5230184. [Accessed: 26-Oct-2021].

11. "Ontario's COVID-19 vaccination plan," COVID. [Online]. Available: https://covid-19.ontario.ca/ontarios-covid-19-vaccination-plan. [Accessed: 26-Oct-2021].

12. "SARS-COV-2 variant classifications and definitions," Centers for Disease Control and Prevention. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html. [Accessed: 26-Oct-2021].

13. "Logistic Regression," Speech and Language Processing. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/5.pdf. [Accessed: 26-Oct-2021].

14."5.7 logistic regression: Interpreting model ... - youtube.com," Youtube. [Online]. Available: https://www.youtube.com/watch?v=tfJUbCGELCQ. [Accessed: 26-Oct-2021].

15. V. Bajaj, N. Gadi, A. P. Spihlman, S. C. Wu, C. H. Choi, and V. R. Moulton, "Aging, immunity, and covid-19: How age influences the host immune response to coronavirus infections?," Frontiers, 01-Jan-1AD. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2020.571416/full. [Accessed: 26-Oct-2021]. l

16. "GGPLOT2 scatter plots : QUICK START GUIDE - R software and Data Visualization," STHDA. [Online]. Available: http://sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization. [Accessed: 26-Oct-2021].

17. C. D. Yihui Xie, “R markdown cookbook,” 10.1 The function knitr::kable(), 07-Oct-2021. [Online]. Available: https://bookdown.org/yihui/rmarkdown-cookbook/kable.html. [Accessed: 26-Oct-2021].

18. Y. Holtz, “R Color Brewer’s palettes,” – the R Graph Gallery. [Online]. Available: https://www.r-graph-gallery.com/38-rcolorbrewers-palettes.html. [Accessed: 26-Oct-2021].